© 2017 by Joseph Ryan Peterson. All rights reserved.

A TALE OF TWO MICROBES: COMPUTATIONAL INVESTIGATIONS OF BIOLOGICAL PROCESSES IN *ESCHERICHIA* AND *METHANOSARCINA* AT MULTIPLE SCALES

BY

JOSEPH RYAN PETERSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Zaida Luthey-Schulten, Chair
Professor Yann R. Chemla
Professor Martin Gruebele
Professor William W. Metcalf

Abstract

Biological phenomena, while grounded in the laws of physics and chemistry, often exhibit behaviours too complex to be ascribed to a single electron density, bond vibration or chemical reaction. Rather, the processes inherent to living organisms—adaptation, growth, homeostasis, metabolism, replication, and response to stimuli—result from concerted function of the physical and chemical interactions of thousands to trillions of molecules interacting with one another. In principle, the governing equations for a biological system could be written in terms of the fundamental physical and chemical laws; however, such an approach would suffer from the intractability in its solution and the incomprehensibility of the result; it would be difficult to see the forest from the trees. This leaves the possibility of study using either analytical studies of toy models or computation.

Adopting the latter approach in this work, a subset of cellular processes including metabolism, growth, and response to stimuli are examined. A variety of modeling approaches are employed to capture phenomena at different scales. Using as model systems *Methanosarcina acetivorans*, an anaerobic methane producing archaeum adapted to niche environments, and *Escherichia coli*, a faculative anaerobic bacteria with diverse capabilities, the influence of extrinsic (environmental) and intrinsic (inherent) factors on the organisms' behaviours is examined.

In Part I of the thesis, studies of *Methanosarcina* species are presented. A kinetic model of methanogenesis—the metabolic pathways unique to Archaea

that produce methane—in *M. acetivorans* is developed and used to examine the sensitivity of methane production rates to abundances of methanogenesis proteins. Subsequently, the metabolism of *M. acetivorans* when grown on several substrates is examined using genome-scale metabolic modeling. Metabolic phenotypes, wherein the methanogens utilize metabolic pathways to different extents, were predicted by integrating RNA expression and half-life data with the models. Strikingly, it was shown that the organism adjusts RNA half-lives of nearly half of metabolic genes to optimize metabolic flux for different growth substrates. This discovery was the first to show such a global role for half-life in defining metabolic phenotype. Concomitantly, the metabolic model was corrected and expanded, especially in the context of the cell's compositional requirements, by adding new terms to the model's biomass equation. Two comparative genomic studies were subsequently undertaken, enabled primarily by the availability of this and other highly curated metabolic models. First, the genomes of all fully sequenced Archaea were mapped across the available metabolic models to examine conservation of metabolic function. This revealed that amino acid metabolic pathways relatively more highly conserved than coenzyme, lipid, nitrogen, and transport metabolism. Second, the metabolic models of several *Methanosarcina* species were mapped across the genomes of 30 *Methanosarcina* species, enabling a pan-reactome study of these metabolically diverse methanogens. By examining the resulting core-reactome in the context the conserved genome, knowledge gaps in the metabolism could be filled. Importantly, by examining the pan- and core-genome of the *Methanosarcina*, a biosynthetic pathway

for methanophenazine, a methanogenesis cofactor, was hypothesized.

In Part II of the thesis, causes of stochasticity and heterogeneity were examined in the model organism *E. coli*. Adopting a simulation technique designed to sample the chemical master equation noise in gene expression was examined. Inspired by the inability of traditional models (which neglected genome replication) to fit the distributions obtained using single-molecular fluorescence in situ hybridization experiments, the effect of genome replication on the noise observed in genes placed at different locations around the circular genome was examined. Simulation results indicated that relaxation of the RNA count from a pre- to a post-replication steady-state significantly affected the shape of the resulting distributions. Analytical results showed that the noise of a constitutively expressed gene could be completely defined by three variables: the location of the gene on the chromosome, the RNA half-life, and the cell doubling time. Overall, this showed that previous studies that neglected to handle genome replication explicitly could both qualitatively and quantitatively misinterpret experimental data. Finally, adopting a completely different modeling framework, metabolic cooperativity in *E. coli* colonies growing on agar surfaces were examined. Building upon previous work that identified acetate cross-feeding using reaction-diffusion partial-differential equations, the effects of strain specific differences in the metabolic capacities and geometrical confinement were examined. The behavior of five different *E. coli* strains were examined; it was found that the extent and timing of metabolic cross-feeding were significantly different, even for closely related strains. Finally, cross-feeding

was found to only vary when the growth substrate had abrupt changes in geometry (e.g. a wall or pit), and that smooth changes caused imperceptible changes in growth.

To the American taxpayer.

Acknowledgements

Numerous people deserve thanks for supporting me throughout the last five years. First and foremost, I thank my family—especially my mother and father—for their support, encouragement, patience, and most of all their love throughout the many trials. I thank Jesse Daniel Peterson for teaching me what it means to be an honourable man and for keeping me humble at times when my ego liked to run away.

I thank Professor Zaida Luthey-Schulten for her support throughout my Doctoral studies. My studies would not have been successful without her guidance, direction and support. I am deeply indebted for her for her tireless advocacy. Without her I would not have been able to see the world and meet so many inspiring people, including, to name just a few, Professor Rudolf Thauer (Max Planck Institute for Terrestrial Microbiology), Professor Arieh Warshel (University of Southern California), and Professor Hermann Gaub (Ludwig-Maximilians University of Munich). I thank her for giving me the opportunity to Lattice Microbes at the many workshops throughout the years. I thank her for allowing me the independence in my work that allowed me to thrive. And above all, I am extremely grateful that she taught me how to write, for writing is really the key to success.

I thank my fellow research lab members who made appearances throughout the years for their advice and insight. It was a special pleasure to work with John Cole, Michael Hallock, Lars Kholer, Piyush Labhsetwar, and Tyler Earnest, though I greatly appreciate the camaraderie of all the other mem-

bers, including David Bianchi, Marian Breuer, Ke Chen, Zhaleh Ghaemi, Jonathan Lai, Marcelo Cardoso Dos Reis Melo, and Seth Thor.

The work would not have been possible without the close collaboration of a number of individuals. I thank Professor Jingyi Fei (University of Chicago) and Professor Taekjip Ha (Johns Hopkins University), Dr. Petra R.A. Kohler, Professor Thomas Kuhlman (UIUC), Professor William W. Metcalf (UIUC) for working with me. I especially thank Dr. Matthew N. Benedict for the crash course in methanogens.

I would also like to thank Professors Yann Chemla, Martin Gruebele, and William W. Metcalf for agreeing to advise me as committee members and for providing their insight and expertise in biology and biophysics. Special thanks is extended to Professor Chemla for agreeing to join my committee on such short notice due to the unforeseen medical emergency of another committee member. I thank Professor Gruebele for his advocacy and support in business ventures that have evolved out of the research. And I would be remiss if I failed to thank Professor Metcalf for the many hours and great insight about methanogens he bestowed.

Many national programs supported my work, for which I am thankful. The National Science Foundation Graduate Research Fellowship Program generously enabled much of my research. The Department of Energy supported much of my work on methanogens, and special gratitude is extended to the Office of Science, Biological and Environmental Research for monetary support, and the Joint Genome Institute for providing a plethora of sequencing through the Community Science Program. I thank the National

Institutes of Health Molecular Biophysics Training grant for support during my first year of studies and the National Aeronautics and Space Administration's Astrobiology Institute for their recent support of my work. I thank the National Center for Supercomputing Applications and the Extreme Science and Engineering Discovery Environment for providing computer time that enabled many of these studies. At the end of the day, I thank the American taxpayers for supporting my Doctoral studies through their hard work and sacrifice.

Special thanks to Jacob Faucheaux, Misha Salim, Laura Pepple, Tyler Harpole, and Chris Daly for helping me keep my sanity. Finally, I thank all my friends and partners not yet mentioned for their companionship and love through these years.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Complexity in biology arises due to interactions between multitudes of distinct molecular components. Be it gene expression, metabolism or signalling, nearly every cellular process involves at interaction of more than two components. Be it the universal process of gene expression—wherein transcription factors bind to specific DNA sequences to recruit the RNA polymerase that forms the messenger RNA which must bind to a ribosome to be translated to a final peptide—or the biosynthesis of proline from α-ketoglutarate by-way-of four chemical intermediates, each of which is synthesized by enzymes comprised of multiple subunits, the concerted efforts of numerous molecular machines must work efficiently to counter the forces of entropy. At the cellular or organismal level, the picture is even more complex. Experimental biology, biochemistry and biophysics have uncovered countless details and unprecedented understanding of how life works. However, current experimental approaches are limited to examining just a few interactions with high spatial and temporal resolution. Hence, numerous computational techniques have been developed to examine different scales, interactions and phenomena in biological systems, yielding the aptly-named field of systems biology have arisen.

In this dissertation I will apply a number of these methods to elucidate the behaviour, function and form of biological processes in single-celled organ-

isms at different temporal and spatial scales. Methods from across the field of systems biology will are applied, including: 1) kinetic and steady-state modeling of metabolic pathways, 2) stochastic simulations of heterogeneity among populations, 3) spatial simulations of microbial communities, 4) bioinformatic analyses of gene expression and regulation, and 5) comparative genomics of closely related organisms. The behaviour of two organisms are examined due to their relatively different life styles. The first is *Methanosarcina acetivorans*, an anaerobic methane producing archaeum adapted to niche environments. The second is *Escherichia coli*, a faculative anaerobic bacteria with diverse capabilities. Broadly speaking, the influence of extrinsic (environmental) and intrinsic (inherent) factors on the organisms' behaviours is examined.

The remainder of this chapter will introduce various information necessary to understand the broader context of the work. First, methaogens, their capabilities and justification for their study is presented. Second, *E. coli* is briefly discussed. Third, the importance of stochasticity in biology is discussed in the context of gene expression and population heterogeneity. Fourth, a cursory description of the various modeling techniques is presented. Fifth, the contributions of the various researchers to each of the chapters is acknowledged. Finally, an overview of the remaining chapters is presented.

1.1 Methanogens

Methanogens, as their name implies, produce methane. These strictly anaerobic organisms, which are members of the Archaea domain of life [1], were originally studied due to their relative exotic lifestyles [2]. Originally isolated from hydrothermal vents [3], they are now known to exist in nearly every terrestrial environment from cold [4] to hot [3], the bottom of the ocean [5], to the inside of humans and cows [6]. Seven orders of methanogens are now recognized, which range in their metabolic capabilities [7].

They are broadly interesting for a number of reasons. As one of the largest biological sources of methane—producing on the order of a billion tons a year [5]—they are interesting to environmental and climate scientists [8]. Due to their unique capability of producing methane, they have captured the interest of engineers around the world, who are attempting to harness them as a source of renewable fuel [9] and to optimize them for degradation of sewage [10]. From a scientific perspective, they are model Archaea from the Euryarchaeota branch. And until very recently, the largest sequenced Archaea was a methanogen, capable of methanogenesis from a diverse subset of molecules in a variety of environments, necessitating complex regulatory networks.

Broadly, two types of methanogens exist, class I and class II [11] (though some have proposed three types [12]). Class I are generally capable of hydrogenotrophic methanogenesis, fixing carbon dioxide via:

$$CO_2 + 4\,H_2 \longrightarrow CH_4 + 2\,H_2O$$

These organisms must live in an environment with a source of hydrogen, either abiotic (*e.g.* from a hydrothermal vent) or biotic (*e.g.* from sulfur-reducing bacteria). Type II methanogens as a whole are capable of several additional forms of methanogenesis [13], including aceticlastic methanogenesis, wherein acetate is split:

$$CH_3CO_2H \longrightarrow CH_4 + CO_2,$$

methylotrophic methanogenesis, wherein a C-1 compounds is bifurcated within methanogenesis, parts being oxidized to derive the reducing equivalents to reduce the other fraction:

$$4\,CH_3OH \longrightarrow 3\,CH_4 + CO_2 + 2\,H_2O,$$
$$4\,CH_3NH_2 + 2\,H_2O \longrightarrow 3\,CH_4 + CO_2 + 4\,NH_3,$$
$$2\,(CH_3)_2NH + 2\,H_2O \longrightarrow 3\,CH_4 + CO_2 + 2\,NH_3,$$
$$4\,(CH_3)_3N + 6\,H_2O \longrightarrow 9\,CH_4 + 3\,CO_2 + 4\,NH_3,$$
$$2\,(CH_3)_2S + 2\,H_2O \longrightarrow 3\,CH_4 + CO_2 + 2\,H_2S,$$
$$4\,CH_3SH + 2\,H_2O \longrightarrow 3\,CH_4 + CO_2 + 4\,H_2S,$$

and methyl reduction methanogenesis, wherein hydrogen provides the reducing equivalents for methyl-group reduction:

$$CH_3OH + H_2 \longrightarrow CH_4 + H_2O,$$
$$CH_3SH + H_2 \longrightarrow CH_4 + H_2O.$$

These pathways are shown in aggregate in Figure 1.1. Different type II methanogen species have different sets of methanogenesis capabilities. Methanogens of the genus *Methanosarcina* have the largest repertoire characterized to date including growth on methanol, tri-/di-/mono-methylamine,

dimethylsulfide, methane thiol, methyl-mercaptopropionate, and H_2/CO_2. Therefore, they are perhaps the most interesting methanogens to study.

Largely due to this fact, I focused on *M. acetivorans* [14], which is capable of nearly all forms of growth but for that on H_2/CO_2. The plethora of growth substrates, and the biosynthetic pathways have been thoroughly characterized [15–17, 17–25], and require complex regulatory networks to tune gene expression in such a way to allow flux to flow through the methanogenesis pathways in the correct directions [17, 26, 27]. In addition, they can be genetically modified [28, 29], there is sufficient data was available to perform computation, including a genome-scale metabolic model [30–32], and transcriptomic/proteomic data, much of which was generated during our investigations [33–36]. During my thesis research, I generated kinetic model of methanogenesis which was capable of modeling the growth on methanol, trimethylamine and acetate, and a minimal transcriptional regulatory network [35] (the focus of **Chapter 2**). Curious about regulation of metabolism, I subsequently integrated gene expression measurements and half-life data for growth on methanol, trimethylamine and acetate with a genome-scale metabolic model to predict differential pathway usage [36] (the focus of **Chapter 3**). Finally, expanding beyond *Methanosarcina* the genome-scale metabolic models were propagated to other Archaea [37] and members of the *Methanosarcina* genus to allow examination of the pan-reactome and metabolic diversity of these organisms (the focus of **Chapters 4 & 5**).

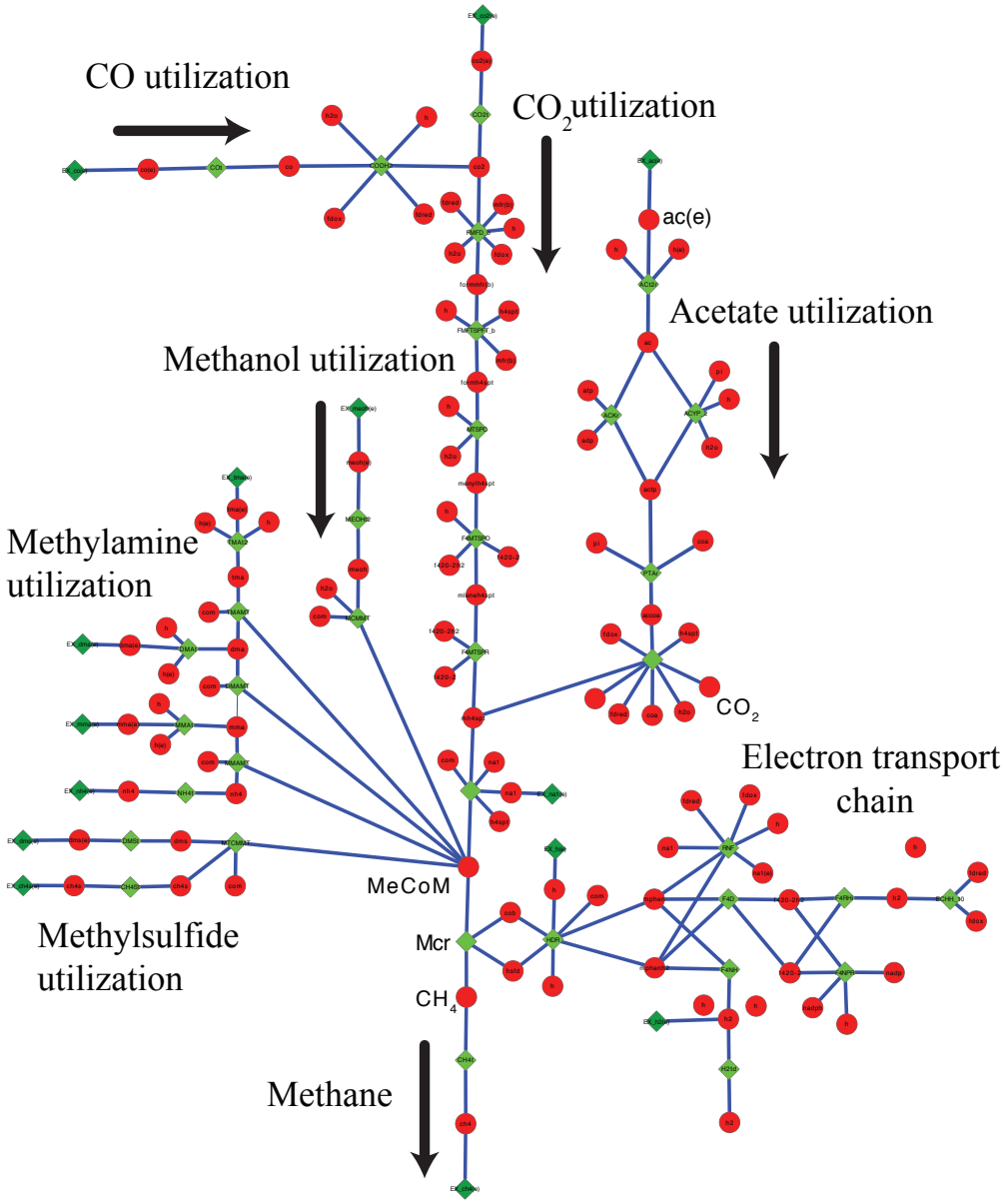


Figure 1.1: **Methanogenesis Map.** Map of the methanogenesis pathways from *M. acetivorans*. The entrypoints for the various growth substrates are indicated.

1.2 *Escherichia coli*

E. coli is a household name primarily due to the numerous outbreaks of pathogenic strains that are highly publicized in the media. These gram-negative, facultative anaerobic bacteria are the most widely studied microbe in the world, rivalled perhaps only by yeast. *E. coli* is capable of growth on numerous sugars, can ferment during anaerobic growth, and are adapted to be generalists. For these reasons they contrast with the specialist *M. acetivorans*, providing a different model system and the second microbe that is studied during my thesis research. Investigations into heterogeneity in *E. coli* are described in **Chapters 6 & 7**. Due to the wealth of literature describing *E. coli*, much of which will more elegantly describe them than possible here, I will go no further.

1.3 Stochasticity in Biology

Stochasticity is a key consideration when modeling single cells or populations, as deviations in a single feature, *e.g.* the count of transcription factor molecules, between two cells result in drastically different behavior. Apparent randomness in cellular processes is magnified by the generally low count of cellular constituents (*i.e.* RNAs, DNAs, proteins, transcription factors, *etc.*). For example, a change by one unit in the number of a transcription factor that is on average found to be in five copies in a cell constitutes a 20% change in the effective strength of activation or repression (ideally). Stochasticity in

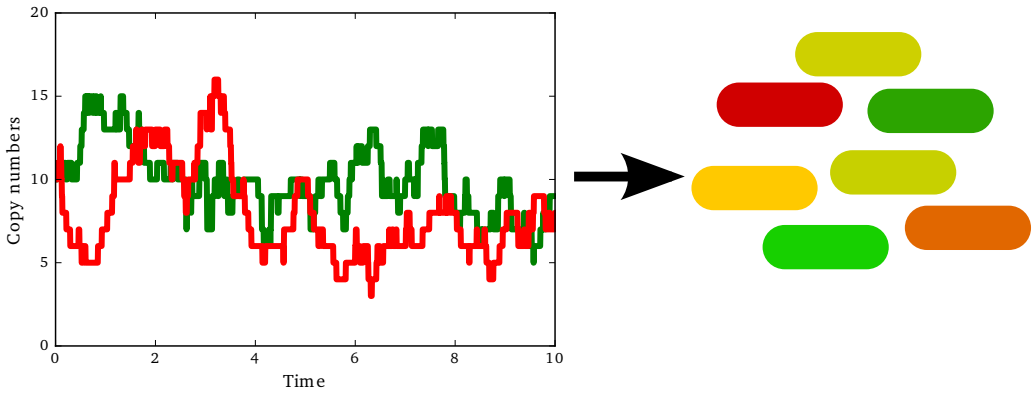


Figure 1.2: **Stochasticity in Single Cells.** Two genes with identical transcription, degradation and translation rates may yield differences in protein copy numbers due to stochasticity in the underlying chemical processes. If, for example, the genes encoded two differently colored fluorescent proteins, cells within a population would express a spectrum of colors.

gene expression was first demonstrated by Swain *et al.* in 2002 in a brilliant experiment [38]. They placed a cyan fluorescent protein and yellow fluorescent protein at equal distances from the origin of replication under the same promoter and measured fluorescences of each cell (see Figure 1.2). These experiments demonstrated that individual cells in a uniform environment can be in different states. It is generally said that stochasticity gives rise to "noise" within a population.

Stochasticity is inherent in every chemical process within a cell and results in different levels of noise. Two types of noise are generally considered: intrinsic noise, that inherent with the actual chemical process, and extrinsic noise, that arising from everything else [38]. For example, in stochastic gene expression, the random transcription of a gene gives rise to intrinsic noise, while variability in transcription factor, RNAP or ribosome copy number,

gives rise to extrinsic noise. Studies throughout the years have examined simple gene expression [38–41], regulated gene expression [42–46], how feed-forward and feedback loops in regulatory networks effect expression [47,48], and more recently how genome replication effects gene expression [49,50] and correlations among sources of noise shape the proteome landscape [51]. Ever more sophisticated treatments of stochasticity are being explored, with recent work showing how it can give rise to spatial heterogeneity in cell division [52], localizations of mRNAs [43] and ribosomal intermediates [53, 54], and how cell architecture affects rates of chemical processes [55].

Ultimately noise is important as it gives rise to population heterogeneity. As examples, populations of *E. coli* and yeast exhibit distributions of metabolic pathway due to differences in protein expression [56,57]. Ultimately, such differences can give rise to long range spatial heterogeneity in colonies of cells [58,59]. This is, perhaps, one of the most important directions of research in stochasticity.

1.4 Modeling Techniques

Numerous modeling techniques have been employed in the simulation of biological processes. At the level of single-cells and communities of cells the most common modeling approaches include mass-action ordinary differential equations, reaction-diffusion partial differential equations [60], and genome-scale metabolic modeling [61] if cell-to-cell variability is neglected. When capturing stochastic effects, common approaches include chemical

master equations [62], reaction-diffusion master equations [62], or agent-based modeling [63]. This is not an exhaustive list, especially in light of the ever increasing list of hybrid methods developed to capture biological phenomena in ever expanding detail.

Here, I will describe brief four approaches utilized in this thesis. First, mass-action kinetics will be discussed, as it is used in solving the kinetics of methanogenesis in **Chapter 2** [35]. Second, genome-scale metabolic modeling and the closely related flux balance analysis will be described, as it is used in **Chapters 3, 4, & 5** [36, 37]. Third, stochastic approaches that can be used in examining cell-to-cell variability in gene expression as applied in **Chapter 6** is discussed [50]. Finally, a hybrid reaction-diffusion/genome-scale modeling approach used to examine spatial heterogeneity in *E. coli* colonies is discussed as it pertains to **Chapter 7** [59].

1.4.1 Mass-Action Kinetic Approaches

Mass-action kinetics describe the dependence of the time-rate of change of a system of reacting chemical species on the current state of the system. The system is modeled as a set of non-linear ordinary differential equations (ODE):

$$\frac{d\vec{C}}{dt} = \mathbf{K}\mathbf{S}\vec{C}^{\mathbf{S}} \tag{1.1}$$

where $\vec{C}$ are the concentrations of the chemical species and $\mathbf{K}$ is a matrix encoding the rates of reactions and $\mathbf{S}$ encodes the stoichiometric dependencies

for each reaction. In the case where all reactions are elementary reactions, $\mathbf{K}$ is a matrix of constants. For example, a chemical system:

$$\mathrm{A} + 2\,\mathrm{B} \underset{\mathrm{k}_r}{\overset{\mathrm{k}_f}{\rightleftharpoons}} \mathrm{C}$$

can be written as:

$$\frac{d[A]}{dt} = -k_f[A][B]^2 + k_r[C] \tag{1.2}$$

$$\frac{d[B]}{dt} = -2k_f[A][B]^2 + k_r[C] \tag{1.3}$$

$$\frac{d[C]}{dt} = -\frac{d[A]}{dt} \tag{1.4}$$

Often in modeling biological systems, the elementary rates are unknown and model equations must be used (and $\mathbf{K}$ becomes a function of certain chemical concentrations). For example in enzyme kinetics, the Michaelis-Menten equation is used:

$$\frac{d[P]}{dt} = \frac{k[E][S]}{K_m + [S]} \tag{1.5}$$

where $[E]$ is the enzyme concentration, $[S]$ is the substrate concentration, $[P]$ is the concentration of product, and K_m is the eponymous Michaelis constant. Various alternatives to these equations exist that capture inhibition, competition, *etc.*. And in cases where more complex chemical interactions are modeled, for instance cooperative binding of multiple ligands to an enzyme or the concerted efforts of several transcription factors to the activity of a gene, phenomenological expressions such as the Hill equation are used. For

example, for cooperative activation the differential equation becomes:

$$\frac{d[P]}{dt} = k\frac{[S]^n}{K^n + [S]^2}, \tag{1.6}$$

where for repression the differential equation becomes:

$$\frac{d[P]}{dt} = k\frac{K^n}{K^n + [S]^n} \tag{1.7}$$

Once written, these systems of equations are trivial to numerically solve using any of a plethora of ODE solvers. In addition to describing the time-dependent behavior of a biological system, the equations can be used to examine the sensitivity of the system. Sensitivity analyses characterize the response of the system to a perturbation in one of the rates or initial conditions. While sophisticated methods have been developed to analytically characterize the sensitivity, it is trivial to compute numerically for all but the largest systems. One good measure is the relative local sensitivity:

$$s_i = \frac{X_i}{Y}\frac{\partial Y}{\partial X_i}\bigg|_{X_0 \neq X_i} \tag{1.8}$$

where Y is the response and X_i is one of the independent variables (rates or initial concentrations). This measure was employed in **Chapter 2** to examine how methanogenesis flux depended concentrations of each enzyme.

1.4.2 Genome-Scale Metabolic Modeling Approaches

Kinetic modeling is powerful in its predictive capabilities; unfortunately, in modeling biological systems such as metabolism, many of the rates are

unknown rendering the system intractable with such a kinetic description. Although several heroic efforts have recently generated genome-scale kinetic models for *E. coli* [64, 65], these required integration of numerous transcriptomics, proteomics and metabolomics datasets which are unavailable for most organisms. The paucity of kinetic data proved to be a significant hurdle in the 1980s, 1990s and 2000s during a time when systematic constructions of metabolism were being compiled [66–68], prompting the development of methods applicable to genome-scale modeling.

Most notable and successful among these methods is flux balance analysis (FBA) [69]. FBA is a steady-state approach that has its roots in methods used in chemical engineering. Analogously to a chemical factory, FBA makes the assumption that on the timescales of interest, flux in and out of any reaction (pipe) is balanced, and solves for the distribution of fluxes through metabolism (the chemical plant) based on several assumptions. The problem boils down to solving the equation:

$$\mathbf{S} \cdot \vec{v} = 0 \tag{1.9}$$

where $\mathbf{S}$ is the stoichiometric matrix and $\vec{v}$ is the set of fluxes through the metabolic reactions. Limits are generally imposed on the fluxes:

$$\vec{a} < \vec{v} < \vec{b} \tag{1.10}$$

to capture constraints on metabolic fluxes (*i.e.* enzyme capacity constraints, uptake and secretion limits, *etc.*). Posed in such a way, the system of equa-

tions can be solved using linear programming techniques. The system is under-determined, therefore an "objective" must be specified. Generally, in prokaryotic systems, the objective is taken to be maximizing biomass production (which is generally considered to be analogous to growth rate, though need not necessarily be the same). To do this, a biomass reaction is added to the system, usually of the form:

$$S_{RNA}\,[RNA] + S_{DNA}[DNA] + S_{Protein}[Protein] + ... \xrightarrow{v_biomass} [Biomass]$$

where equivalents of each of the cell building blocks $[X]$ are converted to a unit of biomass in stoichiometric amounts S_X. Typical building blocks include DNA, RNA, amino acids, lipids, vitamins and cofactors, and sometimes osmolytes. By specifying the objective, the linear program can be solved and an optimal solution found. The solution is not unique so conventionally a parsimonious solution, one that minimizes the total flux, is selected using parsimonious FBA [70].

The outputs of an FBA simulation are the growth rate of the cell and the distribution of fluxes through metabolic pathways. Typical units of fluxes are mmoles of substrate per gram dry cell weight per hour (mmol/gDCW/hr). While FBA is a powerful tool at predicting growth rates within the regimes for which the models were calibrated, there is a desire to use them to predict other scenarios. As such, various methods developed to integrate data from transcriptomics, metabolomics, experiments, etc. have been developed [71–74]. A new method in the spirit of some of these enhanced FBA methods was developed and employed to study the effect of gene regulation on pathway usage in **Chapter 4**.

1.4.3 Stochastic Approaches

To capture the heterogeneity in cell populations originating from the stochastic processes discussed previously, more sophisticated approaches than those previously discussed must be employed. Indeed, while in general the deterministic ODE solutions to reacting chemical systems described above capture the right means, they are completely incapable of describing the heterogeneity, and in some scenarios are even incapable of capturing the average behavior. The most common approach to model cellular processes such as gene expression the master equation is employed [62]. The master equation is a first-order ordinary differential equation:

$$\frac{d\vec{P}}{dt} = \mathbf{A}\vec{P} \tag{1.11}$$

where $\vec{P}$ is a vector describes the probability of the system to be in each state i and $\mathbf{A}$ is the transition matrix. Generally, for gene expression, $\mathbf{A}$ is taken to be time-independent, therefore the equation describes a Markov jump process.

For spatially homogeneous systems, the system is described by the chemical master equation (CME) where $\mathbf{A}$ is a set of rate constants that are proportional to their macroscopic analogs (*e.g.* the rate measured in a well-stirred reaction flask). The CME can thus be written in a more verbose form:

$$\frac{dP(\vec{s})}{dt} = \sum_{r}^{N_r} a_r(\vec{s} - \mathbf{S_r})P(\vec{s} - \mathbf{S_r}) - a_r(\vec{s})P(\vec{s}) \tag{1.12}$$

where the first summation runs over all potential reactions in the system

(N_r), $\vec{s}$ describes the state of the system (*e.g.* counts of each chemical species), the first (second) term describes the flux of probability into (out of) a state $\vec{s}$, $\mathbf{S_r}$ describes the stoichiometry of the reaction r, and the a_r is the propensity of the reaction to occur. The propensity function a_r connects the stochastic description rate to the macroscopic rate constants, *e.g.*:

$$
\begin{aligned}
a_0(x) &= k_0 x N_A V \delta t && (1.13)\\
a_1(x) &= k_1 x \delta t && (1.14)\\
a_2(x,y) &= k_2 \frac{xy}{N_A V} \delta t && (1.15)\\
a_3(x,y,z) &= k_2 \frac{xyz}{(N_A V)^2} \delta t && (1.16)\\
&\ldots && (1.17)
\end{aligned}
$$

where k_i is the ith order macroscopic rate constant, V is the volume of the reaction system, N_A is Avogadro's number, and δt is a time intervale.

The CME is directly solvable only for the simplest of scenarios, such as constitutive gene expression [40–42]. Indeed, during the work in **Chapter 6** an analytic solution to gene constitutive gene expression where genome replication is explicitly handled is presented [50]. While in principle the CME is solvable numerically integrable via matrix exponentiation

$$P(\vec{s},t) = Ce^{\mathbf{A}\int P(\vec{s})dt} \tag{1.18}$$

this is numerically intractable due to the large state space. Approximate

methods such as the finite state project that truncate the state space have been created [75], but these still suffer the curse of dimensionality. A much more common solution is to simulate many realizations of chemical systems that follow the underlying physics. The most common solution is the stochastic simulation algorithm (SSA) of Gillespie [76]. The SSA is employed in the study of gene expression in **Chapter 6**.

When spatial heterogeneity needs to be captured, the more general reaction-diffusion master equation (RDME) may be employed. In the RDME the master equation transition matrix now describes both reaction and diffusion:

$$\mathbf{A} = \mathbf{R} + \mathbf{D}. \tag{1.19}$$

Again simulation is the most commonly used method to get at the evolution of the system. Many solutions have been proposed, with a notable version from the Luthey-Schulten laboratory encapsulated in the Lattice Microbes software package [77,78]. While I have simulated the RDME in some projects, none of them feature in this thesis and I will not dwell further on the method.

1.4.4 A Hybrid Reaction-Diffusion/Steady-State Method

As mentioned previously, countless hybrid approaches have been developed to study biological phenomena. The approach used in **Chapter 7** of this thesis is that of Cole *et al.* [58]. The method couples reaction-diffusion partial differential equations (PDE) with genome-scale metabolic modeling. The

method discretizes space into a regular cubic lattice. Chemical concentrations are modeled via a reaction-diffusion PDE:

$$\frac{\partial \vec{C}}{\partial t} = \mathbf{D}\lambda^2 \vec{C} + R(\vec{C}) \tag{1.20}$$

where $\vec{C}$ is a vector containing the concentrations of metabolites, $\mathbf{D}$ encodes the diffusion rates of the metabolites and $R(\vec{C})$ encode the reactive fluxes of the species. R includes any reactions among chemical species, active and passive transport into and out of cell volume and, crucially, exchange fluxes computed via a local dynamic FBA (dFBA) [79] simulation (more precisely, fluxes are read from a table of solutions computed via FBA and the solution is used to compute uptake and efflux). The reaction-diffusion equation is solved on a 3 dimensional regular cubic lattice, generally via a central finite difference scheme [80].

Cells, while being represented as a volume fraction (ϕ_i) on the lattice, are not diffused or actively transported (*e.g.* via chemotaxis) among lattice points. Rather, as cell growth occurs they are pushed isotropically into neighbouring lattice points (after some maximum volume fraction within the lattice site is achieved, namely $\sum_i \phi_i \geq 0.65$). Volume fraction is related the mass of cells as

$$\phi_i = \frac{m_i}{V \rho_i} \tag{1.21}$$

where m_i is the mass of a particular cell type in the lattice site, V is the volume of the lattice site, and $\rho_i = m_{i,cell}/V_{cell}$ is the density of a single cell.

Cell mass grows exponentially at the rate set by the local dFBA as

$$\frac{dm_i}{dt} = v_{bm,i} m_i \tag{1.22}$$

where v_{bm} is the flux through the biomass equation. An absorbing boundary condition for the cell mass is applied to the boundaries of the simulation volume. Cell mass is prevented from penetrating into the agar substrate. The reaction term in Eq. 7.1 is coupled to the cell mass and the predicted uptake flux as $R(\vec{C}) = \vec{m} \cdot \vec{v}_C$ where $\vec{v}_C < \vec{v}_{C,max}$. The maximal uptake/secretion rate, $v_{C,max}$, is constrained assuming enzyme saturation effects (e.g. Michaelis-Menten kinetics for glucose uptake) and to prevent a chemical in a lattice site from becoming negative ($\vec{C} \geq 0$). The cell volume fraction couples to the chemical concentrations by hindering diffusion (*e.g.* an attenuated diffusion rate computed according to a diffusion law that considers the local cell volume fraction [81]) and via the reaction term discussed above. Volume fractions (ϕ_i) for each cell phenotype are tracked at each lattice site, and a "regulation" function allow cells to transition between phenotypic states depending on the local concentrations of the chemical species.

1.5 Contributions to Various Chapters

Needless to say scientific research is no longer performed in a vacuum. Thoughts, ideas and data are now readily available through the internet and the scientific body of knowledge is growing faster than a single person can comprehend. Thus, the scientific process requires more multi-disciplinary

and collaborative efforts than ever. My research is just the same; dozens of people from several different fields have all contributed to this work. While it would be impossible to list every contribution from each collaborator, the projects here include input from: Dr. Mathew N. Benedict, Dr. John A. Cole Jr., Dr. Jeremy R. Ellermeier, Taekjip Ha, Michael J. Hallock, Dr. James R. Henriksen, Dr. Ankur Jain, Dr. Lars Kohler, Dr. Petra R.A. Kohler, Dr. Piyush Labhsetwar, Dr. Judy Luke, Dr. Zaida Luthey-Schulten, Mary-Beth Metcalf, Dr. William W. Metcalf, Dr. Nathan D. Price, ShengShee (Seth) Thor, Dr. Sarah Stevens, Dr. Nicholas Youngblut, and Dr. Rachel Whitaker. Here, I will attempt to give credit for the major contributions where due (aside from those of the PIs as their roles in guidance and oversight of the research should be obvious) and describe my specific contributions to each of the projects presented below.

The work in **Chapter 2** resulted from a collaboration between the Ha, Metcalf and Luthey-Schulten laboratories. Dr. Labhsetwar helped in analysis of the data, preparation of the figures and manuscript, Dr. Jain performed the SimPull experiments, Drs. Ellermeier and Kohler grew the methanogens, performed genetic manipulations and generated sequencing data. I developed the kinetic model, performed growth simulations and analysis, and constructed the draft transcriptional network.

The work in **Chapter 3** resulted from a collaboration between the Metcalf and Luthey-Schulten laboratories. For this work, I am deeply indebted to both Drs. Kohler for growing the methanogens and generating transcriptomics datasets. ShengShee Thor helped modifying the metabolic models

for *M. acetivorans*, performed the arduous task of laying out the metabolic map for the methanogen, and ran a number of model simulations. My contributions include the analysis of the transcriptomics datasets, the discovery the analysis of regulation across metabolism, and metabolic modeling. Specifically, the major finding of the paper that regulation of the half-lives of metabolic genes is key to optimizing the metabolic pathway usage in different growth substrates, was mine.

The work in **Chapter 4** was performed entirely in the Luthey-Schulten laboratory. This work was a close collaboration between ShengShee Thor and myself, each contributing figures and writing to the manuscript. My unique contributions include the cross-species analysis of the conservation of metabolic capabilities.

The work in **Chapter 5** resulted from a collaboration between the Price, Whitaker, Metcalf and Luthey-Schulten laboratories. Drs. Henriksen, Luke, Stevens, Youngblut, and Mary-Beth Metcalf grew the methanogens, performed the genome sequencing, and reconstructed the genome sequences. Dr. Benedict initiated the work, did the first pan-reactome analysis, generated draft models, and made many of the discoveries in this work. I continued the work started by Dr. Benedict by manually curated the models, identifying novel functions among the methanogens, and analyzing the pan-genome. The figures and manuscript were primarily generated by Dr. Benedict and myself.

The work in **Chapter 6** resulted from a collaboration between the Ha and Luthey-Schulten laboratories. Dr. Fei grew the *E. coli* and performed

the FISH experiments. Dr. Cole and I both contributed significantly to the manuscript. While I discovered that neglecting the mRNA relaxation period was the cause of error in previous treatments of DNA replication and performed all the simulations for the work, Dr. Cole undertook all the analytical treatment of the Master equation; the analytical theory is wholly his. Dr. Cole and I both generated figures and wrote the manuscript.

The work in **Chapter 7** was performed entirely in the Luthey-Schulten laboratory. Using code developed by Dr. Cole, I performed and analyzed the simulations, generated all the figures and wrote the manuscript. Dr. Cole consulted during the performance of the research.

1.6 Research Objectives and Dissertation Overview

The overarching objectives of my thesis are: 1) to examine the capabilities, modes and regulation of growth of methanogens of the *Methanosarcina* genus, and 2) to investigate what types of intrinsic and extrinsic features give rise to heterogeneity among populations of *E. coli*. The thesis is split into two parts:

- **Part I** broadly focuses on investigations of methanogens and comprises four chapters:
  - **Chapter 2** describes development and application a kinetic model for the methanogenesis pathways from *M. acetivorans*. This work was published in *Archaea* [35].

  - **Chapter 3** demonstrates and example of how transcriptomics data can be integrated with genome scale metabolic modeling to predict differential pathway usage in response to growth substrate availability. This work was published in *BMC Genomics* [36].
  - **Chapter 4** examines conservation of metabolic capabilities among the Archaea by propagating the available metabolic models over the archaeal tree of life using bioinformatic approaches approaches. This work was published in *Archaea* [37].
  - **Chapter 5** utilizes a more comprehensive pan-genomic approach to define the pan-reactome in the *Methanosarcina* genus, identify novel metabolic functions, and explore specific differences.

- **Part II** examines two specific types of heterogeneity in *E. coli* in separate studies:
  - **Chapter 6** applies stochastic modeling and analytical theory to examine the effects of DNA replication on mRNA noise. This work was published in *Proceedings of the National Academy of the USA* [50].
  - **Chapter 7** demonstrates how a hybrid reaction-diffusion/metabolic modeling approach can be used explore how heterogeneity in macroscopic bacterial colonies depends on strain-specific growth features as well as environmentally imposed constraints. This work was published in *PLoS ONE* [59].

Part I

Kinetic, Regulatory and Genome-Scale Analyses of Methanogens of the *Methanosarcina* Genus

Chapter 2

Towards a Kinetic Model of Methanogenesis

Progress towards a more complete model of the methanogenic archaeum *Methanosarcina acetivorans* is reported. We characterized size distribution of the cells using differential interference contrast (DIC) microscopy, finding them to be ellipsoidal with mean length and width of 2.9 µm and 2.3 µm respectively when grown on methanol, and on average 2.3 µm long and 1.7 µm wide when grown on acetate. We used the single molecule pull down (SiMPull) technique to measure average copy number of the Mcr complex and ribosomes. In creating a model, RNA expression data (RNA-seq) measured for cell cultures grown on acetate and methanol can be used to estimate relative protein production per mole of ATP consumed. A kinetic model for the methanogenesis pathways based on biochemical studies that have been further validated by recent metabolic reconstructions for several related methanogens, is presented. The kinetic model is capable of capturing experimentally observed methane production rates for cell cultures growing on methanol. In this model, twenty-six reactions in the methanogenesis pathways are coupled to a cell mass production reaction

The contents of this chapter are based in part on work previously published as Joseph R. Peterson, Piyush Labhsetwar, Jeremy R. Ellermeier, Petra R.A. Kohler, Taekjip Ha, William W. Metcalf, Zaida Luthey-Schulten. "Towards a Computational Model of a Methane Producing Archaeum," *Archaea*, vol. 2014 Article ID 898453, 18 pages, (2014) [35]. Specifically, P.L. created figures 2.1-2.3, 2.6-2.9 and analyzed the cell characteristics as well as helped craft the paper, A.J. performed the SiMPull experiments, and J.R.E. and P.R.A.K. provided experimental support.

that updates enzyme concentrations. The archaeum's growth was most sensitive to the number of methyl-coenzyme-M reductase (Mcr) and methyl-tetrahydromethanopterin:coenzyme-M methyltransferase (Mtr) proteins. A draft model of transcriptional regulation based on known interactions is proposed which we intend to integrate with the kinetic model to allow dynamic regulation.

2.1 Introduction

Molecular signatures of ribosomal rRNA evolution were used by Carl Woese and his associates to establish the three primary groupings of living organisms: Archaea, Bacteria, and Eucarya [1, 82–86]. Although the ancestral or communal origins of these three domains remains a matter of debate, increasingly large amounts of data regarding the RNA phylogeny and molecular makeup of cells accumulated over the last several decades continue to support the division between the three primary domains [87]. Furthermore, comparative analysis of the sequences of proteins and RNA involved in translation provides strong evidence that the existence of highly developed translational machinery was a necessary condition for the emergence of cells as we know them [88, 89]. Molecular signatures in the ribosome—idiosyncrasies in its rRNA [87] and/or r-proteins characteristic of each domain of life—were locked in place at the time of evolutionary divergence, destined to become molecular fossils. As Woese postulated in his theory of genetic annealing, ancestors of the three primary groupings of organisms

developed into a number of increasingly complex cell types. The various subsystems of the cell "crystallized," i.e., became refractory to lateral gene transfer with the translation apparatus probably crystallizing first.

While the rRNA phylogeny is supported by phylogenetic analysis of concatenated protein sequences of the fundamental genes in the translational machinery, the effects of lateral gene transfer (LGT) among organisms in the three domains of life are clearly seen in the aminoacyl-tRNA synthetases, a modular subsystem that charges tRNA and helped to establish the genetic code [90–92].

As one moves beyond the information processing systems of translation and transcription, an increasing amount of LGT also extends into other cellular networks. If the early community of cells was more like a modern bacterial consortium, the cells could have cross-fed one another not only genetically but also metabolically. Every improvement in translation that increased its accuracy would have permitted new proteins to emerge, which in turn could have further developed the metabolic pathways within cells. With metabolic functions being modular in nature, these genes could be transferred laterally. Many cases are now known in which a bacterial metabolic gene occurs in one or a few Archaea or vice-versa and has prompted the search for signatures in the metabolic networks that are distinctive of the Archaea [93–96].

Many theories of early life argue for a reducing environment in which anaerobic organisms would likely be the first to have evolved [97]. A phylogenetic analysis of proteins that are distinctive of Archaea and its main

subgroups has led to hypotheses in which methanogens—anaerobic archaeal organisms that derive all of their metabolic energy by reduction of single carbon compounds to methane—feature prominently in the early evolution of life. Methanogens are phylogenetically diverse group of strict anaerobes estimated to produce a billion tonnes of methane per year [5]. They are found in niche environments including shallow and deep hydrothermal vents [98], swamps, paddy fields, land fills [99], hot-springs and oxygen-depleted sediments beneath kelp beds [5].

The *Methanosarcineae* are the most metabolically diverse methanogens known. Only *M. acetivorans* and other archaea in the genus *Methanosarcina* use all four known metabolic pathways for methanogenesis under different growth conditions. While systems biology studies have long used *E. coli* as a model organism in understanding the response of cellular networks to changes in various environmental conditions or gene knock-outs, computational models of methanogen metabolism are just beginning to be established [30,32]. Based in part on our own work modeling genetic switches in *E. coli* and the effects of heterogeneity in protein expression on the metabolism of large populations of bacteria [56], we present here our progress toward a comprehensive computational model of a methanogen. In doing so we have been profoundly influenced by our association with Carl Woese, who published the first genome of a methanogen, *Methanocaldococcus janaschii*, and greatly inspired our interest in characterizing both the translational and metabolic machinery of methanogenic Archaea [100].

We have focused our study on *M. acetivorans*, the several reasons: First,

the organism can grow on three classes of substrates demonstrating use of three methanogenic pathways: 1) methylotrophic pathway—wherein the organism grows on methyl containing substrates including methanol, tri-, di- and mono-methylamine (TMA, DMA, MMA), and methylsulfides (DMS, MMS), 2) acetoclastic pathway—wherein the organism grows on acetate, 3) carboxidotrophic pathway—wherein carbon monoxide is oxidized to acetate, formate and methane [14, 15, 101]. Second, the genome of *M. acetivorans* has been sequenced [14], and considerable effort has been expended towards determining the regulation of gene expression of methanogenesis proteins [102]. Third, the genome exhibits considerable homology to two other well studied members of the genus *Methanosarcina*: *M. barkeri* and *M. mazei* and therefore a model for one will likely be easily modified to work for the others.

Developing a model of the archaeum requires characterization of its physical and biochemical properties. To that end the physical dimensions of the cells, including their length and width, were measured. Modeling also requires estimation of protein/ribosome copy numbers in single cells; the single molecule pulldown (SiMPull) technique [103]—a marriage of the conventional pull-down assay with single molecule fluorescence microscopy—was used to measure the mean copy number of two key proteins. The first protein measured was the γ subunit (McrG) of methyl-coenzyme-M reductase (Mcr) complex as a proxy for number of Mcr complexes, which catalyze the methane producing step of methanogenesis. Second, the ribosomal protein Rpl18p in the large subunit of ribosome, was counted as a proxy for the number of ribosomes. A kinetic model for methanogenesis pathways

capable of representing growth on methanol and acetate was developed using RNA-seq data and kinetic parameters from literature. This model captures several features of comparable experimental data [16, 104]. The model further allows us to probe the sensitivity of the growth (and indirectly the methane production) on the copy number of each protein, directing further experimental study. In an effort to extend the model to simulate growth on other substrates, we compile a list of all experimentally known and hypothetical transcriptional regulatory interactions. These interactions will be used to modulate protein expression as a function of growth substrate that we can marry with the kinetic model in future.

2.2 Experimental and Computational Methods

2.2.1 Strains, Media, and Growth Conditions

M. acetivorans C2A strains (wild-type, WWM 889 :: *SNAP-mcrG*, and WWM890 :: *rpl18p-SNAP*) were grown in single cell morphology [105] at 37°C in high-salt (HS) medium containing either 125 mM methanol or 40 mM acetate [106]. Handling and manipulation of all strains was carried out under strict anaerobic conditions in an anaerobic glove box, using sterile anaerobic media and stocks. Solid media plates (HS medium, 1.5 % agar) were used for selection of SNAP integrants in two steps: puromycin (Research Products International, Mt. Prospect, IL) at a final concentration of 2 µg/ml was used for selection of strains carrying puromycin transacetylase (*pac*), and the purine analogue 8-aza-2,6-dia-minopurine (8-ADP) (Sigma, St Louis, MO)

at a final concentration of 20 µg/ml was used for selection against the *hpt* gene [29, 107, 108]. All plates were incubated in an anaerobic intrachamber incubator [109]. Standard methods were used throughout for isolation and manipulation of plasmid DNA from *E. coli*. DNA purification was performed using appropriate kits (OmegaBio-Tek, Norcross, GA). Growth was quantified by measuring the optical density at 600 nm (OD_{600}, Milton Roy Company Spectronic 21 spectrophotometer) and generation times were calculated during exponential growth.

2.2.2 Genetic Constructs in *Methanosarcina acetivorans*

Genetic fusions with SNAP were made by first constructing plasmids with the SNAP gene near an *aphII* cassette flanked by NheI restriction sites. pJK1048A was used as the template for making fusions to the C-terminus of genes of interest, while pJK1047B was used for fusions to the N-terminus. DNA oligonucleotides (IDT, Iowa City, IA) with homology to the template and gene of interest were used to amplify the SNAP-aphII constructs. The Lambda Red method was then used to integrate SNAP aphII construct into specific N- or C-terminal locations [110], selecting for kanamycin resistance. The *mcrG* and *rpl18p* genes are carried on cosmids created during an *M. acetivorans* cosmid library construction previously performed in the Metcalf lab (Zhang and Metcalf, unpublished). The *aphII* allele was then excised from the cosmid by NheI restriction digest, leaving an in-frame SNAP fusion to the gene of interest. The wild type copies of the genes in question were replaced by the SNAP tagged versions using homologous recombination, as

previously described [108].

2.2.3 Cell Morphology from DIC Microscopy

Cell cultures were grown into exponential phase to an OD_{600} of 0.6 and 1 ml of cultures was removed and centrifuged at 14,000g for 5 minutes. The cell pellet obtained was resuspended in 100 µl HS media without resazurin, and the cells were observed using the differential interference contrast (DIC) microscopy technique on a Zeiss LSM700 confocal microscope.

2.2.4 RNA-seq Analysis

M. acetivorans C2A wild type was adapted to methanol and acetate for 33 generations. The total RNA was isolated from early exponential phase cultures (OD_{600} = 0.4) using TRIzole (Invitrogen, Carslbad, CA) and the Zymo Direct-zol RNA MiniPrep kits (Zymo Research, Irvine, CA). The RNA samples were depleted of the 16s- and 23s-rRNA through hybridization to complementary biotinylated oligonucleotides and subsequent removal with streptavidin-magnetic beads (modified from [111]). Construction of cDNA libraries and high throughput sequencing of RNA was carried out by the Roy J. Carver Biotechnology Center at University of Illinois at Urbana Champaign. All measurements were done in triplicate. The Rockhopper [112] bacterial RNA-seq analysis software was used to map RNA reads to the *M. acetivorans* genome using the default parameters with verbose output enabled. Reads per kilobase per million reads (RPKM) values from the three replicates were averaged and used in subsequent analysis.

2.2.5 SiMPull Experiments

The single molecule pulldown, or SiMPull technique [113] was used to determine mean protein counts for two proteins in *M. acetivorans*. Briefly, SiMPull is a microscopy technique wherein a fluorescently labeled protein of interest is "captured" out of cell lysate by an immobilized antibody attached to a passivated microscope slide. In these experiments, the genetic SNAP-tag system (New England Biolabs (NEB), Ipswich, MA) was used for labeling either the N- or C-terminus of each protein studied (Figure 2.1).

Labeled mutants were grown to exponential phase and harvested at an OD_{600} of 0.6. Cell density was estimated using a Petroff-Hausser counting chamber. One milliliter of cell culture was centrifuged at 14,000g for 5 minutes to obtain cell pellets which were subsequently lysed upon re-suspending the cells in 100 µl of the recommended SNAP labeling buffer: 50 mM Tris-Hcl (pH 7.5); 100 mM NaCl; 0.1% Tween 20; 1 mM DTT (NEB) with 1 µg DNAse. The cell lysate was then incubated with AlexaFluor 488 (NEB) at a final concentration of 10 µM at room temperature for one hour. In order to remove free dye, samples were washed three times with SNAP labeling buffer and concentrated using 10K Amicon ultra centrifugal filters. SiMPull analysis was performed as previously described [113].

Microscope slides were coated with polyethylene glycol (PEG) which minimizes non-specific biomolecule adsorption. Surfaces were doped with 2-5% biotin-conjugated PEG during slide preparation. The bait recruiting rabbit-anti-SNAP antibody (NEB) was immobilized onto the surfaces by

successively flowing in NeutrAvidin (4 µM) and a biotin conjugated anti-rabbit antibody (20 nM), as depicted in Figure 2.3a.

Lysate was washed away and the pulled-down proteins were imaged using a prism type Total Internal Reflection Microscopy (TIRF) with excitation at 488 nm. The resulting images (see Figure 2.3b) were analyzed using custom software as described previously [103], to quantify single spots in the field of view of the microscope. Single spot may correspond to more than one fluorophore which can be discerned by the observation of multiple discrete photobleaching steps (as in the case of Rpl18p, Figure 2.9) indicating that results are lower bounds on the actual number of proteins.

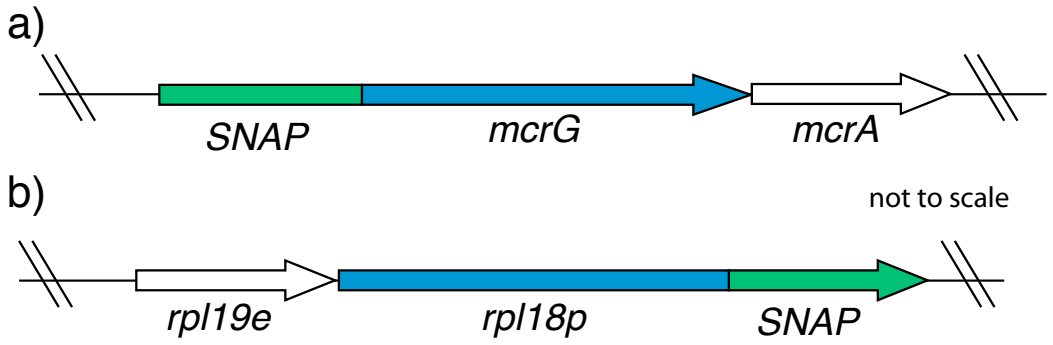


Figure 2.1: **Fluorescently-Tagged Protein Genetic Constructs.** Genetic constructs showing position of SNAP relative to our protein of interest on chromosome of *M. acetivorans*. a) N-terminal label on *mcrG* gene. b) C-terminal label on *rpl18p* gene.

2.2.6 Kinetic Model

RNA-seq expression data for *M. acetivorans* growing on methanol and acetate [27] provide enough parameters for a preliminary kinetic model of the methanogenesis pathways (Table 2.1). The model includes reactions for

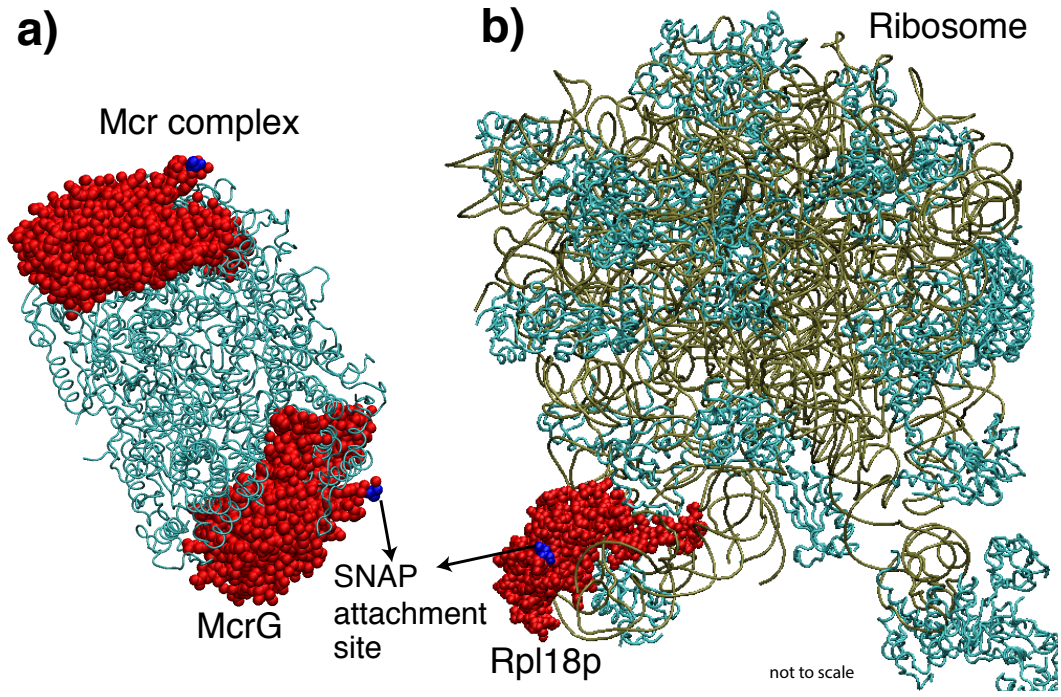


Figure 2.2: **Structures of Fluorescently-tTagged Proteins.** a) Mcr complex from *M. barkeri* (1E6Y [114]) with McrG subunit shown in red with SNAP attachment site shown in blue. b) The Large Subunit of archaeal ribosome (*Haloarcula marismortui*, 4HUB) showing L18p subunit in red and C-terminus where SNAP is attached in blue. These suggest that the position of SNAP is on the outer part of the complexes enabling capture by antibody.

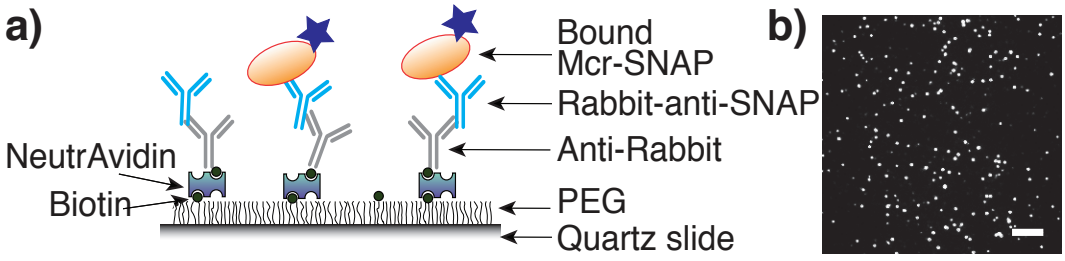


Figure 2.3: **SiMPull Experiments for Protein Count Measurement.** a) Anti-SNAP antibody immobilized on microscope slide using biotin and anti-rabbit antibody, captures SNAP labeled McrG. b) Image obtained where each spot corresponds to at least one immobilized protein.

the methylotrophic, acetoclastic and electron transport pathways shown in Figure 2.4. An additional reaction simulating biomass growth is included to the model that converts ATP created by the methanogenesis driven proton gradient into cell mass. Because 98% of carbons that come into methanogenesis leave as CH_4 or CO_2 [115], ATP is assumed to be a good analog for the growth of the colony. A model schematic is shown in Figure 2.5.

The kinetic model in Table 2.1 is based on the reactions from metabolic model *iMB745* [32]. The reactions are modeled as a set of coupled differential equations (ODEs) which are solved deterministically using the COPASI software [116]. Rate data for 17 of the 26 methanogenesis reactions were taken from the literature [117–131] as reported in the BRENDA database [132]. The other 9 parameters were fit to experiments wherein a cell culture was grown on 125 mM methanol [16]. Three types of reaction mechanisms are used to model the reactions: irreversible unimolecular Michaelis-Menten, irreversible bimolecular Michaelis-Menten and first order. In cases with more than two reactants, the two most important reactants were selected for bimolecular reaction and a constant flux reaction was added that converts the additional reactants to products at the same flux as the rate of bimolecular reaction. In bimolecular reactions, k_{cat} and K_M for both substrates were assumed to be the same. When missing from the literature, K_M parameters for reverse reactions were assumed to be the same as that for forward reactions (e.g. Mtd, Mch, Ftr, Fmd/Fwd). The forward and reverse rate constant are known for Mer giving a ratio of about 6.8. This ratio was assumed for Mtd, Mch, Ftr and Fmd/Fwd as they are in the same pathway. Because Mtr

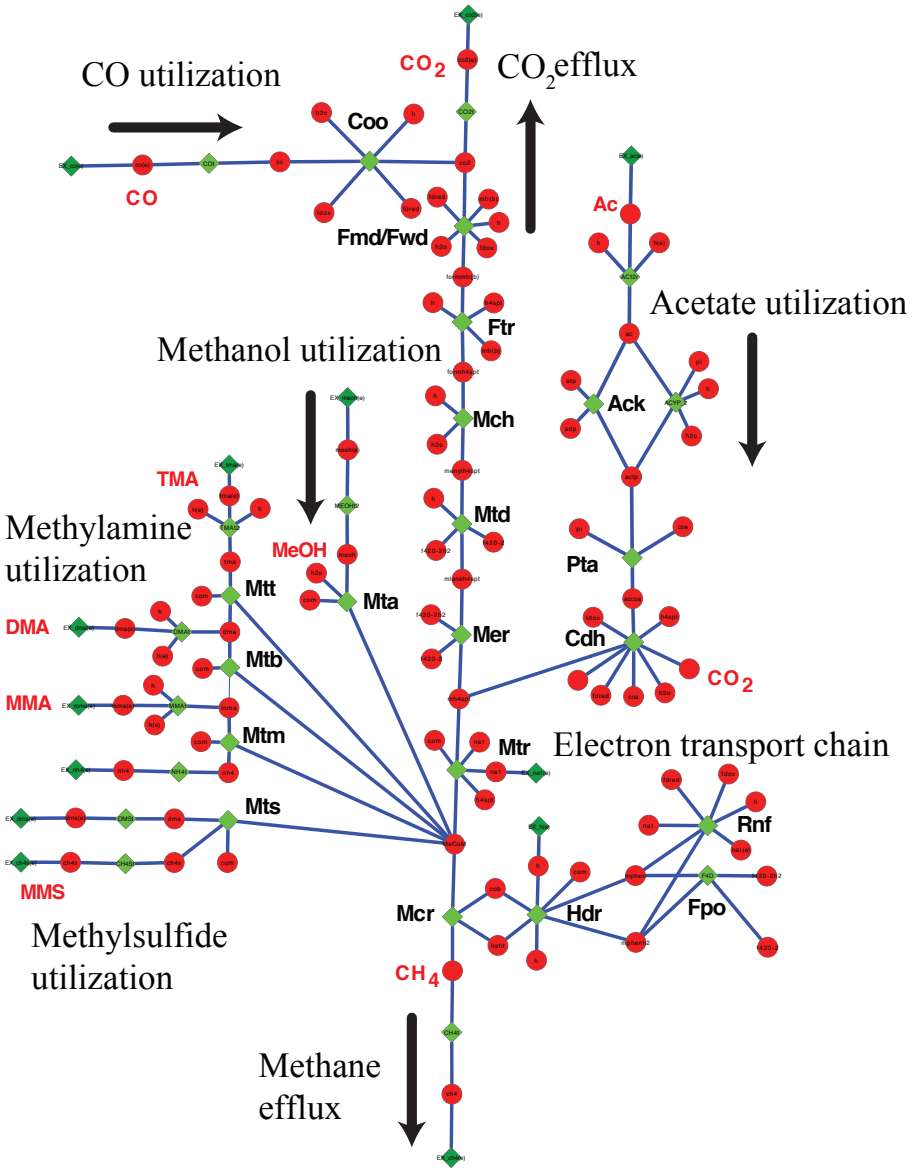


Figure 2.4: **Methanogenesis Pathways.** Pathways from the metabolic map of *M. acetivorans* [32]. Enzymes and metabolites are depicted as nodes while reactions are depicted as edges between these nodes. Enzymes that catalyse reactions are shown as green diamonds and metabolites are shown as red circles. Enzyme names are in black and select metabolite names are in red.

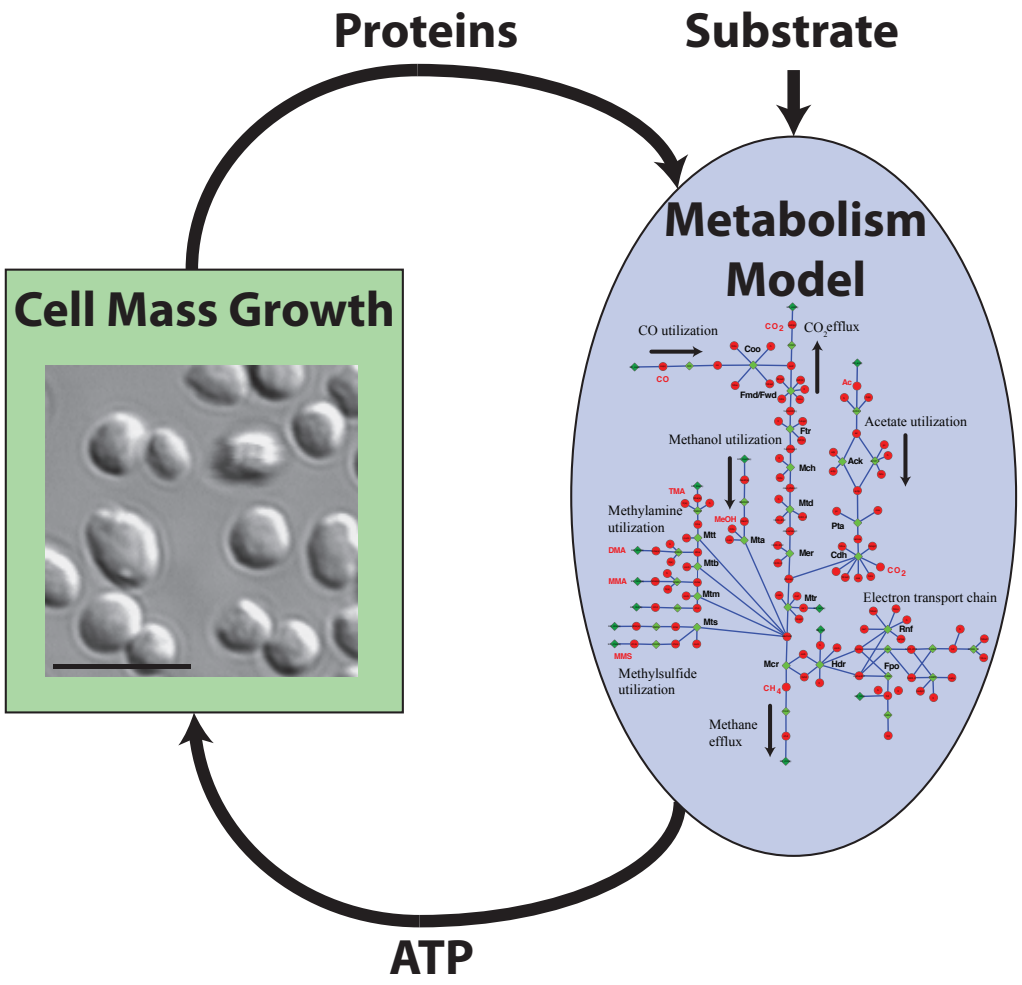


Figure 2.5: **Schematic of the Methanogenesis Kinetic Model.** Flux of ATP from the methanogenesis pathway (See Figure 2.4) feeds into a cell growth term, that updates protein numbers used by the kinetic model, simulating the growth when fed on a certain substrate. The inset in the Cell Mass Growth expression is from DIC microscopy with a 5 µm scale bar.

is known to be nearly at equilibrium [133], we assumed the forward and reverse rates were the same. A value of 50 s^{-1} was chose for this reaction. Finally, Rnf and Fpo were assumed to have similar rates to the Hdr protein as they also catalyze the motion of a similar number of ions across the cell

membrane. The reactions modeled and rate constants used in the model can be found in Table 2.1.

A value of 15.4 grams of cell mass per mole of ATP [32, 135] was used in the biomass expression to match the stationary phase mass of a culture calculated from experimental OD_{420} measurements [104]. The rate of the cell mass reaction was set to match the approximate maximal doubling time of 8 hours known for growth on methanol. The accumulation of biomass in the model leads to an accumulation of enzymes; for each gram of biomass, 63% is assumed to be proteins (in accordance with [32]) of which some are the methanogenic enzymes that themselves catalyze growth. The results of RNA-seq experiments provide estimates for the stoichiometry of methanogenic enzymes per mole ATP. A linear relationship between methanogenic proteins and mRNA was assumed. We determined the relative mass of protein as:

$$0.63M_{total} = \sum_{i=1}^{N_{genes}} a_i \times m_{Protein,i} \tag{2.1}$$

where the coefficients a_i is the mass fraction of ith protein calculated with Equation (2.2). From the value of a_i and the molecular weight of protein $m_{protein,i}$, the number of moles of protein per mole of ATP was determined; these values are provided in Table 2.2.

$$a_i = \frac{m_{protein,i}}{\sum_{i=1}^{N_{genes}} m_{protein,i} \times RPKM_i} \tag{2.2}$$

The model was solved in a 1 ml volume with an initial cell mass of 0.1

mg, calculated from the optical density at the start of growth [16,104]. The concentrations of water and internal protons are assumed to be constant and therefore their effect on the rate constants is implicit and not explicitly modeled. The concentration of extracellular protons was initially set to physiological pH of 7 and are modeled explicitly in the ATP synthase reaction. This reduces the complexity of most of the reactions to either one or two substrate Michaelis–Menten kinetics. Initial concentrations of ATP, ADP, P_i were set to physiological concentrations of 10, 1 and 10 mM [131] respectively. Intermediate energy carriers (CoB, CoM, ferredoxin, etc.) initial concentrations were assumed to be 0.009 mM, which was calculated from the measured value of 474 nmol/gProtein measured for coenzyme F420 in *M. barkeri* grown on methanol [136].

2.2.7 Transcriptional Model

A putative model of transcriptional regulation was constructed using experimental data and inferred regulatory interactions based on gene annotation and sequence homology with proteins known to be regulated in other Archaea. Two different models were developed: the first involving only direct interactions, and the second involving indirect and hypothetical interactions. The direct interactions model was based on experimental evidence of actual binding of the activator/repressor to the promoter region causing up/down regulation of target gene. In addition, genes that showed differential expression and contained the known promoter region characterized as an actual binding site, were included in the direct model. The indirect interaction

model includes interactions reported in the literature where proteins were differentially expressed under different growth conditions, or when expression correlated with a regulator that is differentially expressed, but no direct evidence for the interaction exists. Strength of interactions in the direct and indirect models were taken from the literature; when the transcriptional regulator was overexpressed, the strength of interactions were normalized by the overexpression level. A full enumeration of the literature used to develop these transcriptional regulation models are reported in the Results Section 2.3.5.

2.3 Results and Discussion

2.3.1 Cell Characterization

DIC images of methanol and acetate grown cells were obtained and analyzed in order to quantify their physical dimensions. As seen in Figure 2.7a, DIC microscopy yields enhanced contrast images by taking advantage of a gradient in optical path length between beams of light passing through adjacent points in the illuminated sample. The enhanced contrast is directional, and appears strongest along the shear vector. No contrast occurs perpendicular to the shear vector, which can make the demarcation of cell boundaries difficult. A Hilbert transform has been used in the past with DIC microscopy in order to aid in image segmentation [137]. Custom Matlab scripts were developed to normalize and apply a Hilbert transform to the DIC images. The transformed image shows clearer boundaries around the

imaged cells (Figure 2.7 b). The CellProfiler software was used to identify cell boundaries [138] and measure the cells' dimensions. Figure 2.6 shows the distributions of lengths and widths obtained from approximately 10,000 identified cells. The mean length and width observed were 2.9 µm and 2.3 µm for methanol grown cells, while for acetate grown cell they were 2.3 µm and 1.7 µm respectively. Assuming the cells to be ellipsoid in shape, volume of a methanol grown cell would be approximately 9 fl and that for acetate is approximately 4 fl. Cells have an mean aspect ratios of 1.27 for methanol grown cells and 1.33 for acetate grown cells.

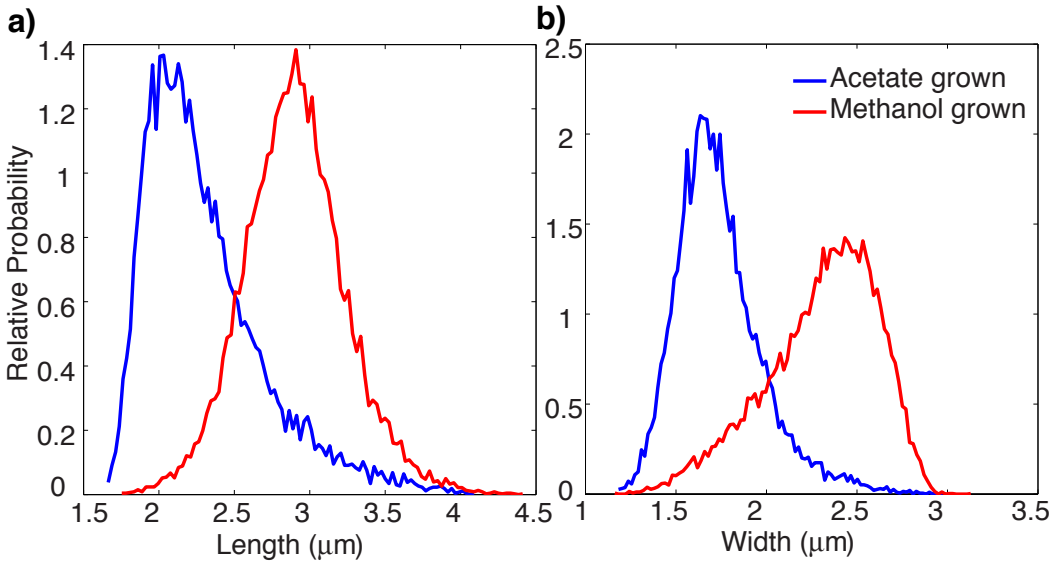


Figure 2.6: *M. acetivorans* **Cell Geometries.** Distributions of a) lengths and b) widths of single *M. acetivorans* cells grown on methanol (red) and acetate (blue) as determined by DIC microscopy and image analysis. Data corresponds to approximately 10,000 cells.

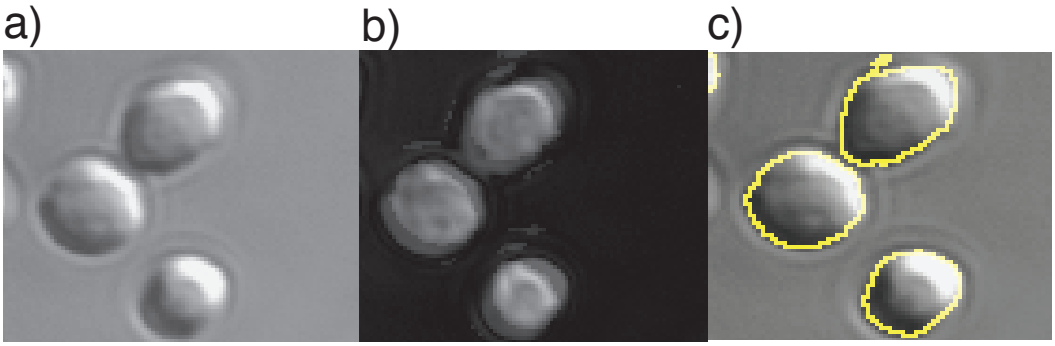


Figure 2.7: **Segmenting** *M. acetivorans* **Cells.** a) DIC images of *Methanosarcina acetivorans* grown on methanol. b) After applying Hilbert transform to DIC images and adding them to the original image, the boundaries of the cells become clearer. c) CellProfiler is used to perform segmentation on images in (b). Here identified cell boundaries are superimposed onto original DIC image to illustrate segmentation.

Table 2.1: **Kinetic Model of Methanogenesis.** Reactions are from *iMB745* [32]. Rate constants used in the kinetic model are taken from literature where indicated or were fit to experiments of growth on methanol. Water molecules and intracellular protons are assumed to be constant and are not explicitly modeled, but are shown in the equations for completeness. The 'Type' column specifies the reaction mechanism: B - irreversible bimolecular Michaelis-Menten, B/C - irreversible bimolecular Michaelis-Menten for the two underlined reactants with a constant flux term for the others, which is set to the flux calculated for bimolecular reaction, M - irreversible unimolecular Michaelis-Menten, F - first order reaction. [a]The reaction that converts ATP into ADP and cell mass, generating proteins via the stoichiometry in Table 2.2. [b]This rate is in units of hr^{-1} which equivalent to an 8 hr doubling time for *M. acetivorans*.

Enzyme	Reaction	k_{cat} (s^{-1})	K_M (mM)	Type	Citation
Acetoclastic Pathway					
Ack	$ATP + Ac \rightarrow AcP + ADP$	1055	0.0713	B	[117]
Ack	$ADP + AcP \rightarrow Ac + ATP$	1260	0.098	B	[118]
Pta	$CoA + AcP \rightarrow AcCoA + Pi$	1500	0.186	B	[119]
Pta	$AcCoA + Pi \rightarrow CoA + AcP$	65.8	0.18	B	[120]
Cdh	$\underline{AcCoA} + 2Fd_{ox} + \underline{H4SPT} + H_2O \rightarrow CO + CoA + 2Fd_{red} + MeH4SPT + 2H^+$	358.5	7.1	B/C	[121]
Cdh	$\underline{CO} + CoA + 2Fd_{red} + \underline{MeH4SPT} + 2H^+ \rightarrow AcCoA + 2Fd_{ox} + H4SPT + H_2O$	1130	0.9	B/C	[122]
Methylotrophic Pathway					
MtaCBA1	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	17	50	M	[123, 134]
MtaCBA2	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	15	50	M	[123, 134]
MtaCBA3	$MeOH + CoM + H^+ \rightarrow MeCoM + H_2O$	5	50	M	[123, 134]
Mer	$MeH4SPT + F_{420} + H^+ \rightarrow MethyleneH4SPT + F_{420}H_2$	119.7	0.25	B	[124]
Mer	$MethyleneH4SPT + F_{420}H_2 \rightarrow MeH4SPT + F_{420} + H^+$	815	0.3	B	[125]
Mtd	$MethyleneH4SPT + F_{420} + 2H^+ \rightarrow MethenylH4SPT + F_{420}H_2$	2650	0.065	B	[126]
Mtd	$MethenylH4SPT + F_{420}H_2 \rightarrow MethyleneH4SPT + F_{420} + 2H^+$	408	0.065	B	–
Mch	$MethenylH4SPT + H_2O \rightarrow FormylH4SPT + H^+$	701	0.57	M	[127]
Mch	$FormylH4SPT + H^+ \rightarrow MethenylH4SPT + H_2O$	100	0.57	M	–
Ftr	$FormylH4SPT + Mfr \rightarrow H4SPT + H^+ + FormylMfr$	1787	0.1	B	[126]
Ftr	$H4SPT + H^+ + FormylMfr \rightarrow FormylH4SPT + Mfr$	262	0.1	B	–
Fmd/Fwd	$FormylMfr + 2Fd_{ox} + H_2O \rightarrow CO_2 + 2Fd_{red} + H^+ + Mfr$	1225	0.02	B/C	[128]
Fmd/Fwd	$CO_2 + 2Fd_{red} + H^+ + Mfr \rightarrow FormylMfr + 2Fd_{ox} + H_2O$	175	0.02	B/C	–
Shared Pathway					
Mtr	$H^+ + MeH4SPT + 2Na_c^+ + CoM \rightarrow H4SPT + 2Na_e^+ + MeCoM$	50	3.7	B	–
Mtr	$H4SPT + 2Na_e^+ + MeCoM \rightarrow H^+ + MeH4SPT + 2Na_c^+ + CoM$	50	3.7	B	–
Mcr	$MeCoM + CoB \rightarrow CoBCoM + CH_4$	5.0	2	B	[129]
Electron Transport Pathway					
HdrDE	$CobCoM + MphenH_2 + 2H^+ \rightarrow Cob + CoM + Mphen + 2H_e^+$	74	0.092	B	[130]
Rnf	$2Fd_{red} + 3Na_c^+ + Mphen + 2H^+ \rightarrow 2Fd_{ox} + 3Na_e^+ + MphenH_2$	80	0.1	B/C	–
Fpo	$F_{420}H_2 + Mphen + H^+ \rightarrow F_{420} + MphenH_2 + 2H_e^+$	80	0.1	B	–
Cell Growth					
ATP synthase	$ADP + P_i + 4H_e^+ \rightarrow ATP + H_2O + 3H^+$	16	0.1	B/C	[131]
Cell Mass[a]	$ATP \rightarrow ADP + P_i + CellMass$	0.125[b]	–	F	[32, 104]

Table 2.2: **Kinetic Model "Biomass Equation".** A list of enzyme stoichiometries in the cell mass reaction. The moles of the indicated protein that are created from 1 mole of ATP calculated as indicated in the Section 2.2.6. [a]Expression values of MtaCBA1 and MtaCBA3 were adjusted such that their ratios to MtaCBA2 were in agreement with the protein expression values measured experimentally [16].

Enzyme	**Methanol (µmol/mol)**	**Acetate (µmol/mol)**	**TMA (µmol/mol)**
Ack	37.5	102.0	36.1
ATP	132.0	406	132.0
Cdh	151.0	134.4	180
Fmd/Fwd	57.4	6.4	13.6
Fpo	27.8	3.96	28.3
Ftr	9.60	4.72	10.3
HdrDE	45.3	38.1	43.4
Mch	30.0	11.6	38.8
Mcr	321.8	398	615.7
Mer	25.5	1.26	36.4
MtaCBA1	1.97[a]	3.57	0.06
MtaCBA2	10.78	5.66	1.57
MtaCBA3	0.20[a]	141.0	0.1
Mtd	36.9	1.69	54.5
Mtr	112.4	144.4	99.3
Pta	36.2	171.0	36.2
Rnf	22.8	17.1	10.0

2.3.2 SiMPull Measurements

SiMPull was used to measure the mean copy numbers of two proteins integral to the growth and physiology of methanogens. The first is the γ subunit of the methyl-coenzyme M-reductase (Mcr) complex. This complex is a lynchpin in the metabolic network, catalyzing the last step of methanogenesis that produces methane. The second is Rpl18P, a ribosomal protein counted as proxy for the ribosome, the protein producing machinery of the cell. We have inserted SNAP genes at the N-terminus of the *mcrG* gene and at the C-terminus of the *rpl18p* gene (Figure 2.1). The fusion of SNAP at the C-terminus of Rpl18P exposes it to the outside of the ribosome enabling the immobilized anti-SNAP proteins to capture whole ribosomes during the SiMPull assay (see Figure 2.2). Using these calibration curves (Figure 2.8), along with estimates of cell density prior to lysing from cell counting experiments, copy numbers of Mcr and ribosomes per cell were obtained (Table 2.3). Mcr numbers agree qualitatively with a recent study where Mcr was imaged on a TEM immunocytochemistry techniques [139].

2.3.3 RNA-seq Experiments

RNA sequencing experiments were performed in order to elucidate the differential expression of methanogenic enzymes on different growth substrates. Comparison of the ratio of mRNA expression on methanol-grown cells to that of acetate-grown cells shows good agreement with the results of quantitative reverse transcriptase polymerase chain reaction (qrtPCR) data

Table 2.3: **Quantification of Cellular Proteins.** Mean protein copy numbers per cells for Mcr complex and ribosomes as estimated by dividing concentration of McrG and Rpl18p subunits in cell lysates by number of cells in the culture. All experiments were done with three technical replicates and two biological replicates grown in methanol.

Methyl-Coenzyme M-Reductase

Biol. Rep.	Conc. (nM)	Cells/ml	Count/cell
1	1.1±.13	$(5\pm2)\times10^8$	1320±713
2	0.37±0.03	$(8\pm3)\times10^8$	273±124

Ribosome

Biol. Rep.	Conc. (nM)	Cells/ml	Count/cell
1	5	$(3\pm1)\times10^8$	10038±3340
2	27.1	$(9\pm3)\times10^8$	18135±6040

previously reported [140]. Two discrepancies are observed, however. The first is that Hdr, which was found to be more highly expressed in methanol-grown cells than in acetate-grown cells, was previously reported to be more highly expressed under acetate growth conditions. This is likely attributable to experimental noise as in both cases the expression ratios are close to 1. The other difference is more pronounced; expression of ATP synthase was found to be 3 times higher on acetate in our experiments, whereas the previously reported results indicate only a two-fold enhancement. In spite of this, expression of methanogenic proteins generally agree with previous reports [27,140], and, importantly, our experiments were run in triplicate and therefore offer greater confidence in our results in addition to some means of error estimation.

Protein numbers, computed using the assumption that they are linearly

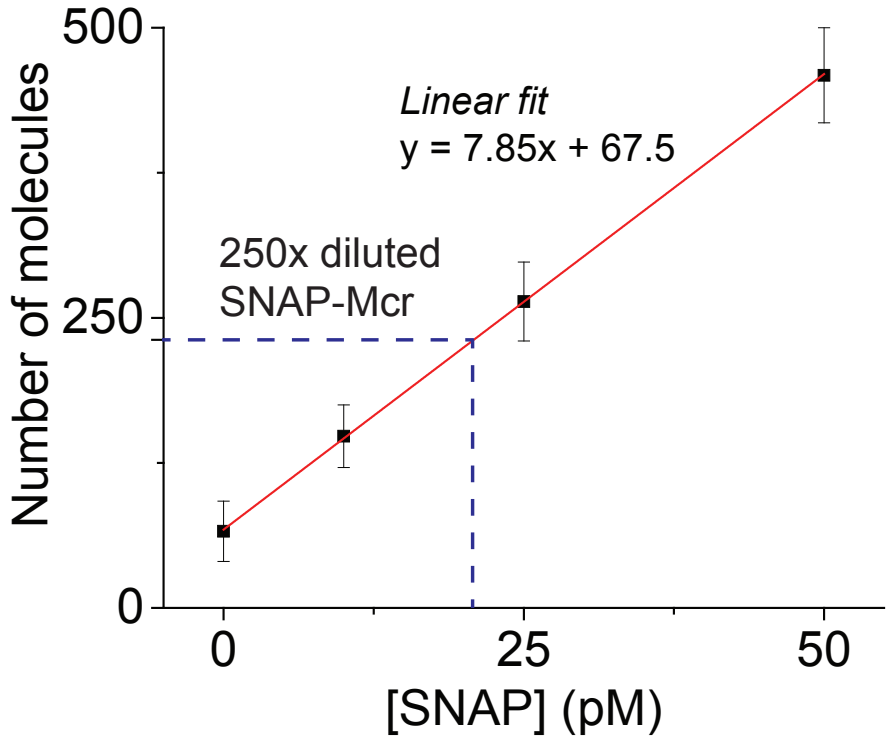


Figure 2.8: **SiMPull Calibration.** Calibration curve for SiMPull experiments that provides a mapping between protein concentration in cell lysate and number of spots observed on the slide.

proportional to mRNA number from RNA-seq, were compared with experimentally measured ones from SiMPull. From [15], we know that OD_{420} of 1 corresponds to 0.41 $\pm$ 0.07 mg dry mass per ml of culture grown on CO. Assuming 63% of this mass being protein and Mcr having 1.2% mass fraction in the proteome, we obtain 3.1 $\pm$ 0.5 µg of Mcr per ml of culture or 10.3 $\pm$ 1.6 picomoles per ml of culture (molecular weight of Mcr = 300 kDa [141]). Using cell density of 500 million cells per ml of culture at OD_{420} of 1 (data not shown), we obtain around 12,400 $\pm$ 2,000 copies of Mcr per cell grown

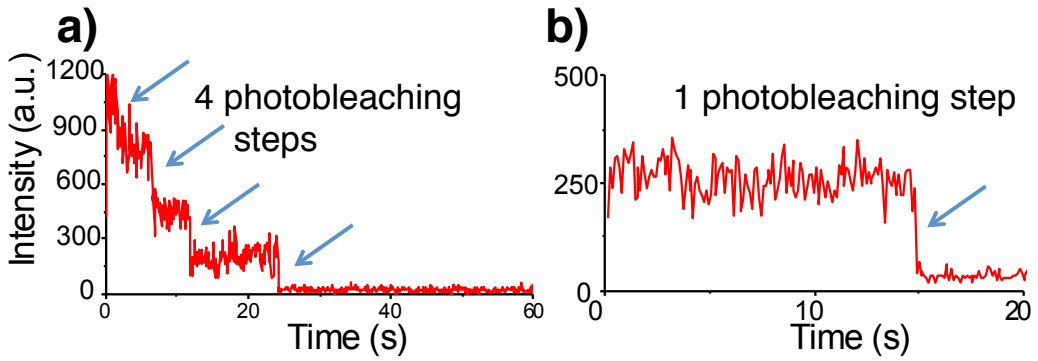


Figure 2.9: **SiMPull Photobleaching Traces.** Multiple photobleaching steps in SiMPull experiments with rpl18p-SNAP strains (a) indicate multiple immobilized proteins as compared to pure SNAP protein which shows only single photobleaching step (b).

in methanol and approximately 6,000 for cells grown on CO estimated due to size differences. SiMPull measurements for Mcr in methanol grown cells ranged from 273 to 1320 copies per cell (see Table 2.3).

2.3.4 Kinetic Model

A kinetic model of the methanogenesis pathways in *M. acetivorans* with single-reaction resolution was developed. The model was fit to cell culture growth experimental results reported previously [16] wherein the methanol consumption rate and OD_{420} of cell culture were studied over time. A comparison of the model fit, seen in Figure 2.10, demonstrates the chosen rate parameters capture the methanol behavior within 10%. The cell mass in the culture was calculated from OD_{420} traces from [16] using the calibration point of 0.41 pm 0.07 mg/mL at OD_{420} 1.0. At maximum growth rate, the model predicts methane formation of 565 nmol/mL×min (slope of simulated curve in figure 2.10) compared with 372±69 nmol/mL×min measured

experimentally [104]. The model correctly predicts the mass of the cells in culture (within 10%), and it captures the 3:1 methane to CO_2 efflux ratio that is necessary for the correct redox intermediate behaviors. Using a stoichiometry of 3.5 protons per ATP, as measured in some experiments for other organisms [142], would make the modeled cell mass growth exactly match the experiments.

As a test of the capabilities of the model to capture methylotrophic growth, it was applied to growth on a mixture of methanol and trimethylamine (see Figure 2.11). The model performs nearly as well as when growing on methanol alone, however the consumption rate of methanol occurs too quickly. This is likely due to the fact the cells are in a regulatory/transcription state to optimally utilize trimethylamine and must change gene expression after trimethylamine is completely utilized. This is reflected in the slowing of biomass growth as trimethylamine is spent (Figure 2.11).

Acetate growth uses a different methanogenesis pathway and is a good test for the rate constants. Using RNA expression values of proteins in acetate-grown cells and the same kinetic parameters obtained from the methanol fit, the results along with experimental results for cells grown in 120 mM acetate [104], shown in Figure 2.12 were obtained. Cell mass entering stationary phase is of the right level, but the rate of growth is much too high. The model predicts a significant build up of carbon monoxide, which should be converted to CO_2 as that step produces more electrons used to drive protons across the membrane. The methane production rate was determined to be 269 nmol/mL$\times$min, which is less than the rate of

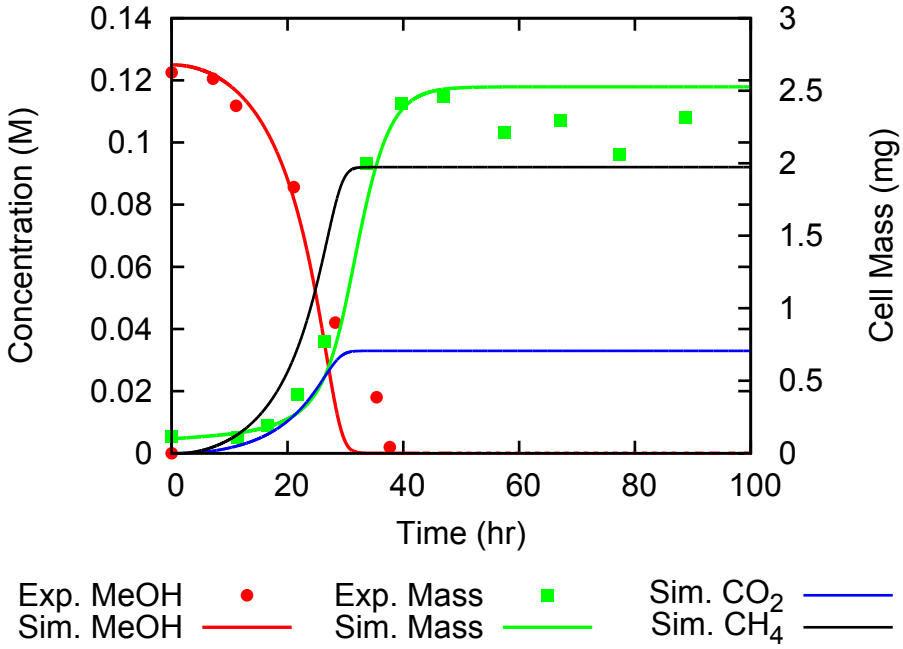


Figure 2.10: **Methanol Growth Results.** Comparison of the kinetic model for growth of *M. acetivorans* culture on 125 mM methanol to the experiment to which it was fit [16]. Lines indicate model results while symbols indicate experimental measurements.

methane production on methanol but is quite a bit higher than experimental measurements of 82 ± 31 nmol/mL$\times$min [104].

This model represents a powerful tool for its ability to be used in testing the sensitivity of cell growth to model parameters such as enzyme copy numbers and rate constants. Moreover, because the growth rate can be thought of as a proxy for the amount of methane produced, understanding its sensitivity to enzyme expression is interesting from a biofuels perspective. The relative sensitivity, s, is calculated using the standard expression:

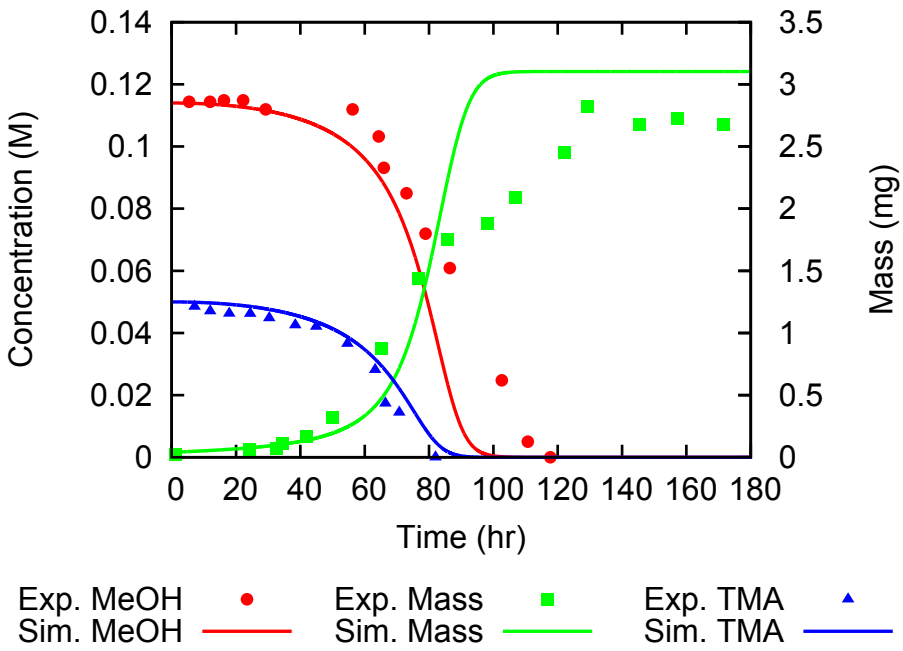


Figure 2.11: **TMA/MeOH Growth Results.** Comparison of the kinetic model for growth of *M. acetivorans* culture on a mixture of 115 mM methanol and 50 mM trimethylamine to the experiment to which it was fit [16]. Lines indicate model results while symbols indicate experimental measurements.

$$s = \frac{x}{Y(x)} \frac{\partial Y(x)}{\partial x} \tag{2.3}$$

where Y is the observable (e.g. growth rate), and x is the parameter (e.g. enzyme copy number). Performing this analysis for methanol growth shows that cellular growth rate is most sensitive to the copy number of Mcr and, in order from greatest to least sensitivity, the copy numbers of Mtr, Rnf, Fpo, MtaCBA2, HdrDE, Mer, and MtaCBA1. This suggests that growth rate is

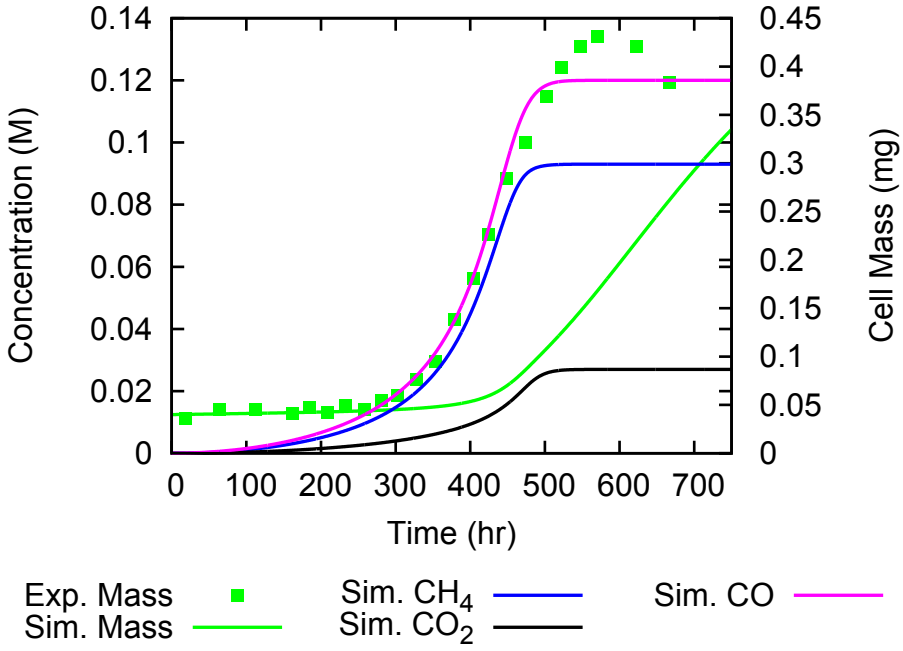


Figure 2.12: **Acetate Growth Results.** Comparison of the kinetic model for growth of *M. acetivorans* culture on 120 mM acetate to the experiment to which it was fit [104]. Lines indicate model results while symbols indicate experimental measurements.

most dependent on the rate at which methyl-coenzyme M can be reduced to methane. These results also indicate that growth rate depends on the equilibrium of species at the branching point that directs substrates either to CO_2 or to methane via the Mtr reaction. This is in line with the fact that the acetoclastic pathway is highly down-regulated during growth on methyl substrates, driving flux through that reaction in the reverse direction. In addition, the rate at which protons are pumped across the membrane and the intermediates regenerated (via Hdr and Rnf) affect the rate significantly. Fi-

nally, it is of no surprise to see that the rate of methane production is strongly dependent on the rate at which methanol is brought into the methanogenesis pathway as demonstrated by the dependence on Mta proteins.

Examination of the sensitivity of growth rate to various enzyme copy numbers under acetate-growth conditions yields a different trend; in order of decreasing sensitivity, Mer, Mcr, HdrDE, Rnf, Mtr. The sensitivity to Mer is directly due to the fact that the reaction can divert flux away from methane production to CO_2 production.

Ongoing work on this model aims to test the behavior on other growth substrates such as CO, MMA, DMA and MMS, as well as mixtures of growth substrates. Future work is planned to refine the rate parameter estimates in order to better capture growth defects with gene knockouts of nonessential methanogenesis genes such as heterodisulfide reductase [104].

2.3.5 Transcriptional Regulation Model

Direct Interactions

The direct interaction map, seen in Figure 2.13a, is largely made up of TATA binding proteins (TBPs) which are common across all Archaea. Other direct interactions were largely identified only for methyltransferases, nitrogen fixation proteins and oxidative stress proteins. Three TATA-box binding proteins (TBPs) were identified in *Methanosarcina* spp. and one experiment characterized their role in regulation [26]. While TBP1 is required for growth, and likely the main transcription regulator, TBP2 and TBP3 are dispensable.

These two differentially regulate approximately 123 genes on acetate versus methanol growth, and the authors of the study concluded that the two transcription factors optimized protein expression for low-energy substrate (e.g. acetate) growth [26]. These interactions are shown in Figure 2.13a.

The second type of direct regulators—those interacting with methyltransferases—act as the mediators for methyl containing organic chemicals entering the methanogenesis pathway. There are separate methyltransferases for each substrate including methanol, trimethylamine, dimethylamine, monomethylamine, and methylsulfide. Because they are some of the most highly expressed genes, they are tightly regulated to preserve the energy balance in the cell [102]. Considerable experimental effort has uncovered eight methyltransferease specific regulators (Msr's). Msr's can act as both up and down regulators. It was found that in the case of MsrA and MsrB, both proteins act in concert to upregulate expression of MtaCB1, and knockout of either can prevent expression [17]. Similarly knockout of either *msrD* or *msrE* will prevent expression of MtaC2 [17]. MsrD and to a lesser extent MsrE also repress MtaC3 [17]. Some Msr's upregulate one gene, while downregulating other genes; for example MsrF enhances expression of methylsulfide methyltransferase *mtsD* on all growth substrates except methanol, while MsrC enhances expression of *mtsF* when on all three methylamines [18,143]. The full set of interactions can be seen in Figure 2.13a.

Nitrogen fixation regulation is the third direct set of regulation interactions found. Two widely conserved nitrogen regulatory proteins named NrpRI and NrpRII have been studied in *Methanosarcinales*. Their regulators

in *M. mazei* Gö1 were found to regulate 23 proteins that shared a regulatory sequence and showed that overall about 5% of all of *M. mazei* Gö1 genes were regulated under nitrogen limitation, with 83 genes up-regulated [144]. Another study showed that several of the unregulated genes under nitrogen limitation were methylamine specific proteins [145]. However support for direct regulation of all 83 genes by Nrp regulators was not established, and these connections are instead included in the indirect interaction map. The method of action for repression was shown to be that NrpRI binds the DNA and NrpRII interacts with the TBPs, preventing the RNAP from binding [146]. Importantly, homologs of NrpRI and NrpRII have been identified in *M. acetivorans* that were differentially expressed under nitrogen limiting versus nitrogen sufficient growth [147], and it is likely that these highly conserved (92-94% identical amino acid sequences) regulators have similar function. An additional two small RNA (sRNA) molecules include Nrp binding sites upstream, $sRNA_{154}$ and $sRNA_{159}$ [148] whose function is as of yet unknown.

One final set of strongly supported interactions is the repression of certain proteins involved in oxidative stress. As Isom et al. point out [149], the MsvR regulator is homologous to a well characterized variant in *Methanothermobacter thermautotrophicus* and 43 genes in addition to the *msvR* gene in *M. acetivorans* contain the two binding sequences upstream of the TATA box. Their study shows support for a homodimer with cysteine residues which likely are oxidized in an oxygen rich environment, causing the dimer to be released from the binding site.

Overall, the total number of interactions in the direct model is 248, with

10 regulators. Strengths of interaction, where known, are indicated by the width of the arrows in Figure 2.13.

Indirect/Hypothesized Interactions

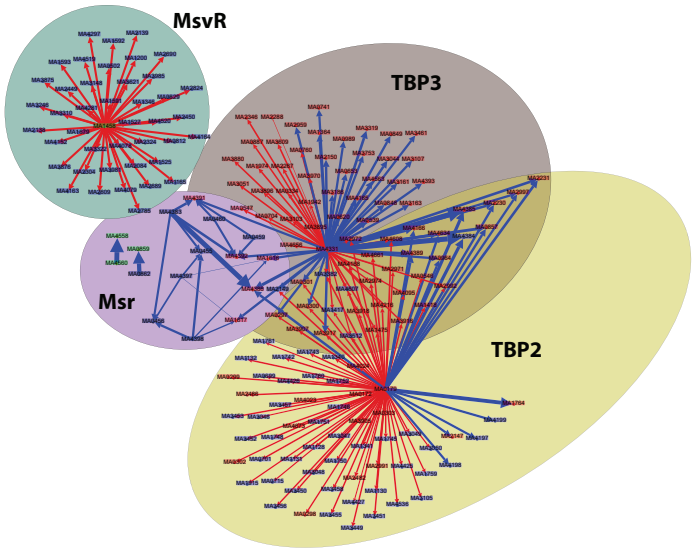
Studies of nitrogen-related regulation in *M. mazei* [144] have led to the identification of 69 proteins that were differentially expressed by at least 3-fold [144,145]. Of the proteins, 35 were involved in nitrogen and energy metabolism, 7 were transport system genes, and 10 were potential regulators. Of particular interest was the up-regulation of the *mtb* and *mtm* genes used in methylamine degradation to generate energy or ammonia, the latter of which must be synthesized from N_2 under starvation conditions. Because many of these genes did not have the binding site for the Nrp regulators upstream of their start sites, they are likely regulated by another protein.

One of the more exciting regulations that has been discovered in Archaea is a sRNA that targets both *cis*- and *trans*-encoding mRNAs called $sRNA_{162}$ [150]. Overexpression of $sRNA_{162}$ in *M. mazei* greatly upregulated many of the methylamine processing proteins. The work also implicated an ArsR family transcription factor as the mediating component in the regulation [150]. Homologs with high sequence identity (>90%) to both the sRNA and the ArsR regulator (gene MA1531) exist in *M. acetivorans*; therefore we have included the same interactions in our hypothetical map.

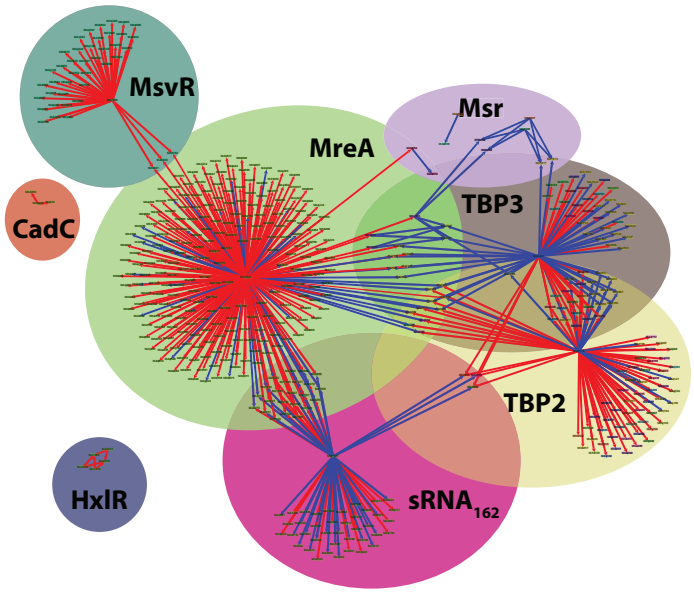
MreA (*Methanosarcina* regulator of energy-converting metabolism) was recently implicated as the global regulator of methanogenesis after it was observed to be 38 fold more highly expressed on acetate than on TMA or

methanol [27]. A study comparing the ratio of methanogenesis protein encoding genes in a strain containing a knockout to the wild type indicated that MreA acts to up-regulate acetoclastic proteins and down regulate methylotrophic pathways [27], the latter of which is mediated by changes in expression of the Msr proteins previously discussed. Therefore, MreA could act as a switch between methanol and acetate utilization. Adding the interactions reported in the paper to the indirect graph allows MreA to have the greatest putative sphere of influence on gene expression (Figure 2.13b).

Cadmium resistance has been studied in a number of different Archaea and bacteria. A well known CadC regulator represses cadmium resistance genes. It is stimulated to unbind by Cd^{2+}, Bi^{3+} and Pb^{2+} [151]. This is a particularly interesting regulation as *M. acetivorans* growing on acetate in the presence of cadmium chloride show between a two- and five-fole increase in methane production, likely attributed to higher levels of acetate kinase and carbonic anhydrase and lower phosphate kinase (Pta) [152]. Furthermore, it has been shown that levels of Coenzyme M increased roughly proportionally to Cd^{2+} concentration [153]. A homolog of the *cadC* gene, MA3940, likely regulates two cadmium efflux encoding genes MA3366 and MA3632 in *M. acetiborans*. In addition, evidence exists for a putative interaction between CadC or one of the genes it regulates and *ack*, *pta*, and carbonic anhydrase. The former two of these interactions are shown in Figure 2.13b.



(a) Direct Interactions



(b) Indirect Interactions

Figure 2.13: *M. acetivorans* **Gene Regulatory Network.** Graph representations of the Direct and Indirect regulations and associated spheres of influence. Red arrows indicate up regulation and blue arrows indicate down regulation. The regulator is indicated by the large black text.

Methanogenesis Gene Regulation

Simplification of the two regulation maps to only those interactions that directly modulate expression of methanogenesis genes yields the map shown in Figure 2.14. From this map it can be seen that two regulators ($sRNA_{162}$ and MreA) interact broadly with methanogenesis genes. Correlations of regulator and methangenesis proteins across several substrates (data not shown) indicate three broad classes of regulation: methanol, acetate, and other methyl-containing substrates (MMA, DMA, TMA, MMS). Regulation motifs usually embody one of two modes of action in many different species: 1) a global regulator that activates/represses many genes, or 2) activation of a single or handful of genes via very specific interaction [154]. Therefore, a possible energetically efficient way to regulate metabolism may be to have a few global regulators that activate/repress many genes, some of which may in turn act as specific regulators capable of fine tuning individual gene expression.

With this knowledge in mind and assuming that the MreA and $sRNA_{162}$ regulators interact in the way proposed in the literature, it can be hypothesized that between the two regulators the three classes of regulation can be covered. In this hypothesis MreA is a global regulator that facilitates the switch between methyl-substrates and acetate. It does this primarily by turning off the CO_2 efflux pathway while turning on the acetate utilization pathway. Upregulation of TMA, DMA, and MMA utilizing proteins is accomplished by expression of the specific regulator $sRNA_{162}$, which turns off expression MA1531, which is hypothesized as a repressor of methylamine

genes.

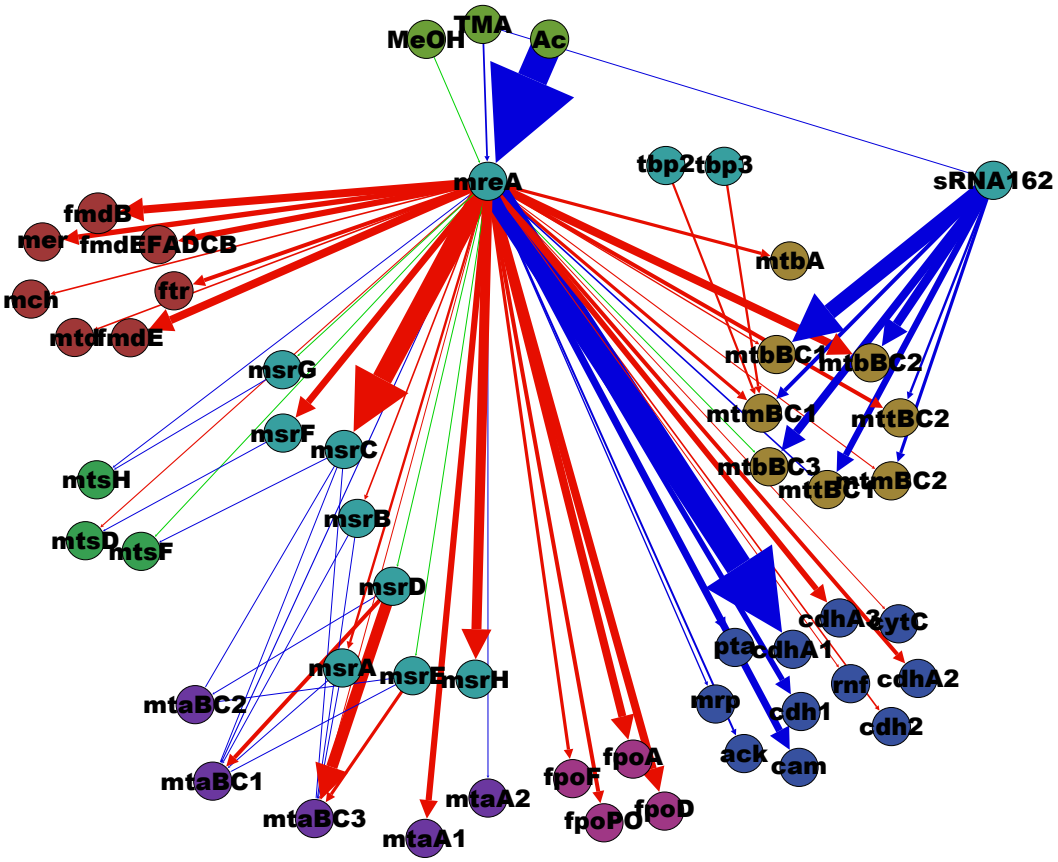


Figure 2.14: **Regulatory Network for Methanogenesis Genes.** A graph representing the regulation of methanogenesis genes. Red arrows indicate down regulation, blue arrows indicate up regulation and green arrows indicate weak regulation. The thickness of the arrow shows the strength of the interaction. Turquoise nodes are known regulators. Red nodes are CO_2 efflux proteins. Olive green nodes are growth substrates. Green nodes are methyl-sulfide methyltransferase proteins. Purple nodes are methanol methyltransferase proteins. Blue nodes are acetate utilization proteins and gold nodes are methylamine methyltransferase proteins.

2.4 Conclusions

We have shown that SiMPull can be used as a method to measure the number of proteins in anaerobic organisms. This entailed genetically engineering a SNAP tag gene into the chromosome along with the gene encoding for the protein of interest, McrG and ribosomal protein Rpl18p. Using this technique we were able to estimate the number of protein complexes in single cells at their exponential phase which is important data for modeling. With Mcr's unique position in the methanogenic pathway, knowledge of its copy number is important for modeling metabolic dynamics. As Ribosomes are known to be one of the dominant components of molecular crowding, their numbers are important to generate accurate *in silico* whole cell models of methanogens.

Using SiMPull and RNA-seq expression data from monoclonal cell cultures of the methanogens growing on acetate and methanol, we were able to estimate the number of proteins in the methanogenesis pathways. Coupling the resulting cell mass growth reaction to the methanogenesis reactions, we were able to fit unknown rate constants to experiments for growth on methanol. Applying the model to growth on acetate, we were able to capture the correct timescale for use of acetate and production of the methane, however the cell mass growth rate in exponential phase was too high.

In order to apply the model to more complex scenarios, especially time varying growth substrate conditions potentially found in the environment, we need a regulation mechanism for expression of the proteins. Examining

correlations of protein expression across different substrates leads to the observation that there appear to be three classes of growth: methanol, acetate and other methyl-substrate (TMA, DMA, MMA, and MMS). Towards the goal of developing a regulation model, we have compiled known transcriptional regulation with putative regulation interactions to create a draft model for *M. acetivorans*. Reducing the draft regulation map to just interactions with methanogenesis protein encoding genes, two regulators arise as global regulators. MreA appears to switch between the acetoclastic pathway and the CO_2 efflux pathway and therefore is hypothesized as the switch between acetoclastic growth and methylotrophic growth. $sRNA_{162}$ appears to turn on expression of genes necessary for utilizing methylamines, and therefore optimizes the organism for methylamine growth.

The physical and stoichiometric properties and kinetic model reported here complement the metabolic reconstructions and constitutes significant progress towards a full computational model of *M. acetivorans*. Spatial heterogeneity, such as that caused by large crowders like the ribosome, is known to cause stochastic effects in similar cells from a monoclonal culture, therefore quantifying the number and distribution is necessary. Because many reactions in methanogenesis occur in the membrane, stochasticity due to the local environment could have a large effect. Larger spatial organization, such as membrane bound protein complex locality and number, can be determined by cryo-electron tomogramography. Such data could be used with the kinetic and regulation models developed here to construct detailed full cell reaction-diffusion models similar to those that have been created previously [43].

Such models would allow study of stochasticity in individual organisms. Ultimately, these models could be used with hybrid reaction-diffusion master equation/flux balance analysis techniques [155] that provide full metabolic modeling with spatial effects due to cell culture organization. The utility of the computational models is that they should be easily extendable to the other *Methansarcina* spp.

Chapter 3

Genome-Wide Gene Expression and RNA Half-Life Measurements allow Predictions of Regulation and Metabolic Behavior in *Methanosarcina acetivorans*

While a few studies on the variations in mRNA expression and half-lives measured under different growth conditions have been used to predict patterns of regulation in bacterial organisms, the extent to which this information can also play a role in defining metabolic phenotypes has yet to be examined systematically. Here we present the first comprehensive study for a model methanogen.

We use expression and half-life data for the methanogen *Methanosarcina acetivorans* growing on fast- and slow-growth substrates to examine the regulation of its genes. Unlike *Escherichia coli* where only small shifts in half-lives were observed, we found that most mRNA have significantly longer half-lives for slow growth on acetate compared to fast growth on methanol or trimethylamine. Interestingly, half-life shifts are not uniform across functional classes of enzymes, suggesting the existence of a selective

The contents of this chapter are based in part on work previously published as Joseph R. Peterson, ShengShee Thor, Lars Kholer, Petra R.A. Kohler, William W. Metcalf and Zaida Luthey-Schulten. "Genome-wide gene expression and RNA half-life measurements allow predictions of regulation and metabolic behavior in *Methanosarcina acetivorans*," *BMC Genomics*, 17(1):924 (2016) [36]. Specifically, L.K. and P.R.A.K. performed the RNAseq experiments supporting this work, S.T. helped with the metabolic modeling and generation of the metabolic map as well as generated figures 3.19-3.23.

stabilization mechanism for mRNAs. Using the transcriptomics data we determined whether transcription or degradation rate controls the change in transcript abundance. Degradation was found to control abundance for about half of the metabolic genes underscoring its role in regulating metabolism. Genes involved in half of the metabolic reactions were found to be differentially expressed among the substrates suggesting the existence of drastically different metabolic phenotypes that extend beyond just the methanogenesis pathways. By integrating expression data with an updated metabolic model of the organism (*i*ST807) significant differences in pathway flux and production of metabolites were predicted for the three growth substrates.

This study provides the first global picture of differential expression and half-lives for a class II methanogen, as well as provides the first evidence in a single organism that drastic genome-wide shifts in RNA half-lives can be modulated by growth substrate. We determined which genes in each metabolic pathway control the flux and classified them as regulated by transcription (e.g. transcription factor) or degradation (e.g. post-transcriptional modification). We found that more than half of genes in metabolism were controlled by degradation. Our results suggest that *M. acetivorans* employs extensive post-transcriptional regulation to optimize key metabolic steps, and more generally that degradation could play a much greater role in optimizing an organism's metabolism than previously thought.

3.1 Introduction

The stability of an RNA molecule, as measured by its half-life, is a critical factor in defining timescales for cellular events. It also sets the correlation time of transient adaptations relative to a cell's growth rate, when compared to its more long-term "basal" phenotype. For example, rapid, high-fidelity responses to extreme external stimuli are mediated through small RNA (sRNA, siRNA, miRNA, etc.) whose function depends largely on search time and efficiency of stimulated co-degradation, sequestration or stabilization when in complex with the target mRNA [156]. On longer timescales degradation finely tunes abundances of critical RNAs [157], controls slow shifts in RNA levels during adaptation between different growth states [158], and contributes significantly to the noise in the steady-state distribution observed in populations of cells [50]. These observations and those of many other studies demonstrate that post-transcriptional control of RNA dynamics is critical to understanding the cellular state; however, the factors defining stability over an organisms entire transcriptome have yet to be fully defined. Furthermore, the consequences of changing RNA stability on metabolic state remains unknown.

Significant strides towards understanding the individual factors affecting RNA stability have been made. To date, genome-wide analyses of RNA stability have been reported for many single-celled, model organisms including representatives of the families *Escherichia* [159–163], *Mycobacterium* [164], *Bacillus* [165, 166], *Sulfolobus* [167–169], *Halobacterium* [169], *Methanocaldo-*

coccus [100, 170] and various yeasts [171–174]. However, the majority of species studied were fast-growing bacterial or eukaryotic species, and archaeal species account for only a small fraction of the whole-transcriptome reports. This study aims to extend our knowledge of RNA stability in archaea by characterizing it in *Methanosarcina acetivorans*, a versatile organism capable of growth and methanogenesis using many substrates and therefore of great importance in the global carbon cycle [5, 14]. The organism has also been implicated in a historical mass extinction as the fossil record shows increase in biogenic methane along with an increase in environmental nickel, an important cofactor in the methanogenesis and the evolution of the acetotrophic (acetate utilization) pathway [175].

Genome-wide RNA stability has been characterized in the first sequenced methanogen *Methanocaldococcus jannaschii* [100, 170]; however, this organism is a type I methanogen only capable of growth wherein electrons derived from hydrogen or formate are used to reduce CO_2 [11]. More complex class II methanogens [11] such as those in the family *Methanosarcinaceae* are capable of growing on a diverse set of substrates including mono-, di-, and tri-methylated molecules as well as acetate, carbon monoxide, and H_2/CO_2; thus, requiring branched methanogenesis pathways and more complex regulation to optimize their growth to a particular environment. They also generally have genome sizes 2-4 times larger than *M. jannaschii*, requiring significantly more regulators, the number of which have been found to scale quadratically in the number of genes [176].

The study of RNA stability in *M. jannaschii* identified noncatalytic cleav-

age sites about 12–16 nucleotides upstream of the translation start site for about a quarter of genes examined, suggesting 5′ leader sequences play a role in post-transcriptional regulation of genes [170]. Several studies posited a similar mechanism could exist in type II methanogens. One study of the operon encoding the acetyl-coenzyme-A decarbonylase/synthase complex in *Methanosarcina acetivorans* measured post-transcriptional regulation to be important in acetotrophic and carboxydotrophic methanogenesis and hints at the possibility that altering transcript stability could play a more global genetic role [23]. A very recent study in a distantly related methanogen *Methanolobus psychrophilus* has demonstrated that both transcriptional and post-transcriptional regulation play important roles providing extra stability in this slow growing, cold-adapted organism [177]. Several studies have discovered small RNAs in the related species *M. mazei* [148, 150]; however, their role in regulating transcript half–lives have yet to be established. Whether post-transcriptional regulation is widespread and whether such regulation is mediated by targeted endonucleolytic degradation or small RNA regulation or translational initiation is yet unknown. Therefore, a characterization of RNA stability in class II methanogens will help us to determine what role degradation plays in the larger context of the cell's economy.

Regulation of gene expression by change in half-life has recently been demonstrated in *L. lactis* and *E. coli* [161–163]. The authors of these papers proposed a method to determine "control coefficients" (which describe whether mRNA abundance is transcriptionally or degradationally controlled) from half-life and expression data. They found that change in growth rate

on glucose manifest small shifts in half-lives and that only about ∼10% of genes were degradationally controlled. To determine the extent to which degradation plays a role regulating gene expression in *M. acetivorans* we performed whole-genome analyses of RNA expression and half-lives in two fast growth substrates (methanol and TMA) and one slow growth substrate (acetate) and applied the control theory. We found, in contrast to the studies in *L. lactis* and *E. coli*, significant shifts in half-life with growth rate and that degradation controls gene expression for up to 28% of genes.

This study provides the first global picture of differential expression and half-lives for a class II methanogen, as well as provides the first evidence in a single organism that drastic genome-wide shifts in RNA half-lives can be modulated by growth substrate. Furthermore, we demonstrate how combining half-lives with expression data can be used to predict transcription rates and average mRNA copies per cell which can in turn be used with computational modeling to predict metabolic phenotype. In the process, we updated the most recent genome-scale metabolic model for *M. acetivorans* to include newly characterized reactions. We used expression data to constrain metabolic fluxes to generate several hypotheses about changes in the metabolic state and metabolite production due to carbon source. We created a metabolic map onto which all information generated in the study could be displayed including reaction energies, enzyme commission numbers, metabolic subsystem, cluster of orthologous group categories, differential expression statistics, half-lives, regulation control coefficients and steady-state reaction fluxes. This effectively creates a visual database which can be used

to understand regulation and the associated metabolic state.

3.2 Methods

3.2.1 Experimental

Strains, media, and growth conditions

M. acetivorans C2A strain (wild-type, WWM82 [29]) was grown in single cell morphology [105] at 37°C in high-salt (HS) medium containing either 125 mM methanol, 50 mM TMA or 120 mM acetate [28, 178]. Handling and manipulation of all strains was carried out under strict anaerobic conditions in an anaerobic glove box, using sterile anaerobic media and stocks. Growth was quantified by measuring the optical density at 600 nm (OD_{600}, Milton Roy Company Spectronic 21 spectrophotometer).

RNA Isolation Procedure

RNA was isolated as previously reported [35]. Briefly, *M. acetivorans* C2A wild type was adapted to methanol, TMA or acetate for 33 generations. Cells were growth in batch cultures tubes. The total RNA was isolated from mid-exponential phase cultures (OD_{600} = 0.4 for MeOH/TMA, 0.2 for acetate growth cells) using TRIzole (Invitrogen, Carlsbad, CA) and the Zymo Direct-zol RNA MiniPrep kits (Zymo Research, Irvine, CA). Specifically, at mid-exponential phase 2 mL of culture were added to 2 mL TRIzole and 4 mL ethanol was added. Samples were processed with the Zymo

Direct-zol RNA MiniPrep kit and RNA was eluted at 50 muL. The RNA samples were depleted of the 16s- and 23s-rRNA through hybridization to complementary biotinylated oligonucleotides and were subsequently removed with streptavidin-magnetic beads (modified from [179]). Samples were stored at -80$^{\circ}C$. Total RNA concentration and integrity were assessed using a Nanodrop® and Agilent BioAnalyzer, respectively. The $A_{260/280}$ ratios measured by the BioAnalyzer were generally $>$2.0 and Nanodrop indicated between 200 and 800 ng/μL RNA obtained. Triplicate biological cultures were processed for each time-point. An additional two and five cultures were processed from mid-exponential phase for TMA and methanol, respectively.

RNA Transcription Inhibition

In order to estimate half-lives for RNA transcripts, a series of RNAseq experiments were performed at timepoints after halting transcription. Transcription was halted by addition of 1 mL of $\sim$ 80 μM of actinomycin D to cultures growing in exponential phase (OD_{600} = 0.4 for MeOH/TMA, 0.2 for acetate growth cells). At 6 or 7 times after transcription was halted 2 mL of culture was withdrawn (5, 10, 20, 30, 60, 120, and 240 min). RNA was isolated as described above. All half-life experiments were performed with three biological replicates.

Sequencing

Construction of cDNA libraries and high-throughput sequencing of RNA were carried out by the Roy J. Carver Biotechnology Center at University of

Illinois at Urbana-Champaign using an Illumina HiSeq2500. All sequence data generated in this report are available online in the GEO database (accession number GSE77738). See Supplemental Section **Additional Methods and Materials** and Supplemental Table S1 for details. Additionally, three RNAseq datasets were taken from the GEO database accession GSE64349.

3.2.2 Computational

Data Quality Control and Normalization

Quality of the RNA reads in each experiment were assessed using the FastQC tool (http://www.bioinformatics.babraham. ac.uk/projects/fastqc/). Individual reads with systematic sequencer error (blocks of unassignable nucleotides, N) were removed, and then reads were trimmed. The adapter sequence was trimmed, constituting between 6 and 12 bases off the head or tail of the read. In some cases 2 bases were trimmed from the end of the read when the Sanger Phred quality score at that base had a score below 20. Trimmed reads were mapped to the reference genome (accession number NC_003552 [180]) using Rockhopper v2.0.2 [112]. The software was set to map single ended reads strand-specifically to the genome. Only the highest scoring mapping for each read was retained. A minimum seed of one-third the read's length was used and the only reads mapping more than 85% of the bases exactly were kept. Resulting reads were considered for further analysis.

Mapped reads were further processed by normalization and averaging.

Two normalization schemes will be considered: 1) reads are normalized per kilobase (gene length) then per one million reads (referred to as RPKM), and 2) reads are normalized per kilobase (gene length), then averaged across operons (see Section **Operon Regulation**) and finally normalized per one million reads (referred to as ORPKM). After normalization, the triplicate biological replicates were averaged for each O/RPKM and the standard deviation computed. These O/RPKM values were used for subsequent analyses (see Supplemental File "ReadCounts.xlsx" for combined data).

Life-time Fitting and RNA Stability Estimation

For each distinct experimental growth condition RPKM values for genes were normalized so that the superoxide dismutase (*MA1574*, *sodB*), which has been characterized as having one of the longest known half-life in the archaeon *Sulfolobus solfataricus* [167], had a half-life of 2 hours to match that measured in *S. solfataricus*. The degradation of each gene was fit to a first-order decay reaction:

$$R_i = R_{i,0} e^{-k_i t} \tag{3.1}$$

using the Levenberg-Marquardt nonlinear least-squares method in SciPy [181]. Here $R_{i,0}$ is taken to be the RPKM for the gene i at time 0 and k_i is the decay rate. The half-life τ_i is then calculated:

$$\tau_i = \frac{ln\,(2)}{k_i} \tag{3.2}$$

Standard fitting residuals were used to compute p-values for the fits as well as for statistical significance testing of half-lives between and within growth conditions. Genes with uncertainty in the fitting value τ_i of greater than 50% were omitted from subsequent analysis. These half-lives can be found in the Supplemental File "HalfLives.xlsx".

RNA structural stability (folding energy) was estimated for each gene by folding the open-reading frame, as annotated in the genome NCBI genome NC_003552, using the RNAFold package [182] using both the Turner 2004 [183] and Andronescu 2007 parameters [184] leaving all other parameters as their default value.

Differential Expression Calling

Differentially expressed genes (DEG) were computed using three statistical models: edgeR v3.8.5 [185], PoissonSeq v1.1.2 [186], and DESeq2 v1.6.3 [187]. For all methods the default normalization provided by the packages was used in computing the DE statistics. A description of the workflow for each of the statistical models can be found in Supplemental Section **Additional Methods and Materials** and code to reproduce the results can be found in Supplementary File "DEGComputation.zip". Genes with a p-value ≤ 0.01 were considered to be differentially expressed. Differentially

expressed genes can be found in the Supplemental File "DifferentiallyExpressedGenes.MultiFactor.xlsx".

Operon Regulation

Four sets of putative polycistronic operon structures for the genome of *M. acetivorans* were considered. One set was predicted by Rockhopper [112], and three sets were taken from the online databases: Microbes Online [188], ProOpDB [189], and DOOR2 [190]. Mapped reads were pooled across operons and differentially expressed analysis was performed (as computed in Section **Differential Expression Calling** and described in Section **Differentially Expressed Genes**).

Computing Evolutionary Conservation

The Integrated Toolkit for Exploration of Microbial Pan-genomes (ITEP) was used to compare conservation of differentially expressed genes [191]. ITEP was used to construct a database of $\sim$60 methanogens with published complete or nearly complete genomes. Default parameters were used to construct the database (E–value cutoff of 1.0 and 1×10^{-5} for nucleic acids and proteins, respectively; MCL clustering was run with the maxbit option with inflation and score cutoffs of 2.0 and 0.4, respectively). For each differentially expressed gene in *M. acetivorans*, the top scoring homolog in each other methanogen was identified, if one exists. These were ordered by degree of conservation and plotted on a phylogenetic tree that is rooted at *Methanopyrus kandleri* using the Python ETE Toolkit [192] (see Fig. 3.11 and

Supplemental Figs. S17 and S18).

Biomass Coefficient Fitting Procedure

A new method was developed to incorporate differential expression data with the metabolic model, as existing methods to integrate expression data into genome-scale metabolic networks have been shown to perform relatively poorly unless metabolomic data was also available and integrated [193]. We reasoned that since mRNA level and protein level generally have poor correlation—in our case, a Pearson r=0.63—that using expression data to make quantitative predictions would be problematic. Therefore, we devised a scheme designed to make qualitative predictions about changes in metabolic flux distributions and the metabolites that the pathways produce. The method takes a single growth substrate as the reference and then varies the biomass coefficients (the required moles of metabolite to create a new gram of dry cell mass) so that the ratio of fluxes between the two predictions matches the ratio of messenger expression for differentially expressed genes. The metabolite requirement in the new condition is then classified as being higher, lower or unchanged with respect to the reference growth substrate, suggesting targets for further experiment.

More specifically the method attempts to minimize an objective function; that is:

$$min\left(\sum_{i=0}^{N_{DE}}\left(\sum_{r|i\in r}\left|\frac{v_{1,r}}{v_{2,r}}-\frac{m_{1,i}}{m_{2,i}}\right|\right)\right) \tag{3.3}$$

for each biomass coefficient b_j. In the equation N_{DE} is the number of differentially expressed genes, $\vec{m}_{c,n}$ is the expression level of gene n in growth condition c, $\vec{v}_{c,r}$ is the flux through reaction r that has a gene-protein-reaction rule that contains gene i in growth condition c. In order to do this, the the list of biomass coefficients is ordered randomly, and the new optimal biomass coefficient $b_{j,opt}$ is found in order from the beginning to the end of the list. For each biomass coefficient, the uptake rate is varied such that the experimentally measured growth rate is achieved. This whole process is performed multiple times with permuted ordering of for optimizing biomass coefficients and the final biomass coefficients are selected as the best for that round of optimization.

The average and standard deviation of these biomass coefficients are computed. If the original biomass coefficient is different from the range of newly sampled biomass coefficients ($p<0.01$, t-test) the biomass coefficient is considered significantly different. If the new coefficient is larger (smaller) it indicates that metabolite is required in higher (lower) amounts than in the reference condition. A large standard deviation indicated that many different selections for that biomass coefficient could give equally good scores (Eq. 3.3).

We used 96 random orderings of the 67 biomass coefficients found in Figures 3.8 and S15. Nine were significantly different comparing TMA (reference condition) to acetate, 12 were significantly different comparing MeOH (reference condition) to acetate, and 16 were significantly different comparing MeOH (reference condition) to TMA.

3.3 Results

3.3.1 Half-Life Distributions

We characterized the half-lives genome-wide for cells growing on acetate, methanol (MeOH) and trimethylamine (TMA) in order to identify changes in transcript "stability" under different growth conditions (see Table S1 and Figs. S1 and S2). Half-lives were extracted from RNA transcriptome profiles measured at six timepoints following transcriptional arrest by the antibiotic actinomycin D as described in the methods. Transcriptome profiles at different timepoints for the same growth substrate were highly correlated (Pearson's r$>$0.94, Fig. S3). The three biological replicates at each timepoint showed a minimum correlation coefficient of 0.99. Both of these observations indicate that our experimental procedure is reproducible (see Supplemental Fig. S3).

Expression profiles were averaged at each timepoint and the time series normalized such that the most stable, core gene found in *sodB* (*MA1574*) [167] had a half-life of $\sim$2 hours. After normalization, most genes had positive half-lives. After fitting the degradation profile to a first-order decay function, 4486, 4487 and 3667 genes in methanol, TMA and acetate grown cultures were found to have positive half-lives with residual error in the fit below 50% of the value of the decay rate (i.e. coefficient of variation, CV$<$50%). Genes with negative half-life or large fitting residual were omitted in subsequent analyses. High-energy yield substrates methanol and TMA had similar

average half-life—59±25 min and 72±29 min, respectively—while the lower-energy yield substrate acetate showed a significantly longer average half-life of 159±59 min (standard deviation, n=3). Probability distributions of the half-lives for each growth substrate where highly statistically different ($p<1.6\times10^{-133}$, Mann-Whitney test) demonstrating that each substrate has unique degradation characteristics. Histograms of the RNA half-lives for the three substrates are shown in Figure 3.1A where the significant shift towards longer half-lives for growth in acetate is apparent; however, a sizable portion of the half-lives remain relatively unchanged as evidenced by the flat profile between 0 and 100 minutes.

In Fig. 3.1B shifts in half-lives for individual transcripts comparing growth conditions and comparing methanogenesis type (methylotrophic vs. acetotrophic) are shown. A "core" set of 1339 genes showed no statistical change among any of the three conditions (t-test, $p>0.01$) which means that $\sim$25% of the genome is not differentially degraded; however, most other genes show some shift in half-life when comparing growth substrates. The observation that not all genes show similar shifts in half-life suggests the organism might employ a mechanism to selectively stabilize/destabilize certain mRNAs. More than 10 times as many genes were destabilized during methylotrophic growth than were stabilized, with a total of 1153 having longer half-lives in acetate than in methanol and TMA.

To further understand the shift in half-lives seen in Fig. 3.1B, genes were binned by clusters of orthologous groups (COGs, the 2014 revision [196]; arCOGs, the 2015 revision [197]). After computing statistics for the distri-

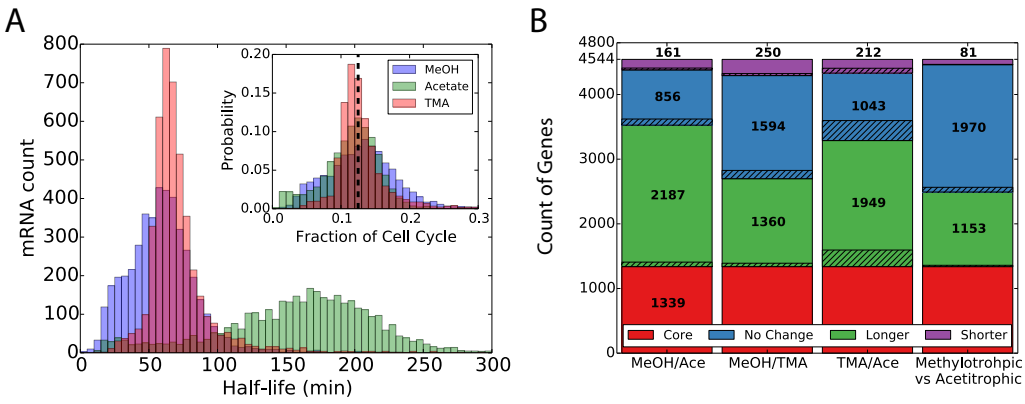


Figure 3.1: **Shift in Half-Life With Growth Substrate.** A) Genome-wide histograms of RNA half-lives for *M. acetivorans* growing in methanol (blue), TMA (red), or acetate (green) media. The shorter lifetimes in high-energy substrates are apparent when compared to the acetate distribution. The inset shows the distribution of half-lives after they have been scaled by doubling time (7.5hr [104, 134, 194], 8.9hr [104] and 24.6hr [104, 194, 195] for growth in MeOH, TMA and Acetate, respectively), demonstrating that the average transcript half-life is a constant fraction of the cell cycle, or about 12.7%±3.5% the doubling time (dashed line). See Fig. S4 for a larger version of the inset. B) A breakdown of changes in half-life by pairwise comparison of growth conditions. Unregulated genes denote that show no statistically significant (t-test, $p>0.01$) shift in half-life under any of the conditions (1339 total; red bar) and those marked as "No Change" (blue bar) do not show significant changes when comparing the indicated conditions. Genes that are stabilized or destabilized when comparing the second condition to the first condition are labelled as "longer" (green) and "shorter" (purple), respectively. Hatched regions indicate the fraction of genes that are differentially expressed in addition to having different half-lives. As discussed more thoroughly in the text, almost half of the stabilized and destabilized genes are common when comparing methylotrophic conditions to acetotrophic growth, suggesting there are similar stabilization mechanism based on either growth rate or substrate.

bution of half-lives in each bin we found that the means many COG classes were significantly shifted between growth substrates (Fig. 3.2 and S5). About 11 classes of genes had significantly longer mean half-life when growing

on acetate than methylotrophic conditions including RNA processing (A), carbohydrate transport and metabolism (G), lipid transport and metabolism (I), cell wall biogenesis (M), cell motility (N), inorganic ion metabolism (P), secondary metabolite metabolism (Q), intracellular trafficking and secretion (U), defence mechanisms (V), extracellular structures (W) and the mobilome (X). Additionally, three classes had significantly shorter half-lives including energy production and conversion (C), amino acid transport and metabolism (E), and translation, ribosome structure and biogenesis (J). It is interesting to note that the three classes that have shorter half-lives in acetate (slow growth) all play major roles in cell growth. Comparing MeOH and TMA growth most classes showed minor shifts except those involved in RNA processing (A), chromatin structure (B), energy production (C), lipid transport (I), replication (K) and extracellular structures (W). The selective stabilization by functional class indicates that *M. acetivorans* uses degradation to prioritize certain functions on different growth substrates.

Only 81 transcripts were stabilized in both methylotrophic conditions compared to acetotrophic growth (see Fig. 3.1B). Using gene set enrichment analysis (GSEA) it was found that translation (ribosomal proteins and initiation factors) and methanogenesis (*mcr*, *cdh*, *mrp*, *hdr*) genes were highly enriched in the transcripts that were particularly stabilized during methylotrophic growth ($p<4.9\times10^{-5}$, $p<2.8\times10^{-2}$, respectively; computed via PANTHER [198]). In general, correlation of half-lives between any two growth conditions was low ($|r| < 0.15$); however, half-lives for genes that were stabilized or destabilized when during methylotrophic growth had a

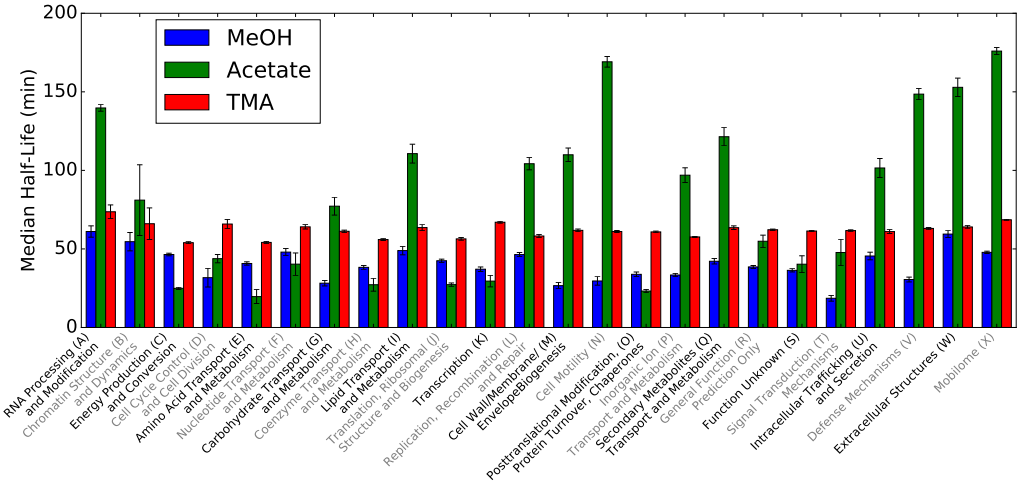


Figure 3.2: **Half-Life Shift by Functional Class.** The median half-lives for the 23 COG classes demonstrate different behaviors for low- and high-energy substrates. The shift in RNA half-lives between substrates are not uniform across functional classes, suggesting there exists a mechanism to selectively stabilize/destabilize the transcripts. See Figure S5 for details about the median and quartiles. Uncertainties were calculated as the weighted standard deviation and are shown as error bars.

significantly higher correlation (r=0.43, $p< 4 \times 10^{-53}$, t-test). This observation might indicate similar stability control for growth in methanol and TMA. GSEA also indicated that transcripts involved with aromatic amino acid synthesis were especially stabilized, encoding growth rate controlled genes in Trp, His, Asp and Phe biosynthesis pathways ($p< 1.4 \times 10^{-2}$; PANTHER).

We found that half-lives did not correlate significantly with RNA Gibbs energy of folding of the ORF as computed using state-of-the-art RNA folding software (r=0.018, p=0.23 by t-test, for both Andronescu2007 and Turner2004 parameters; data not shown). This confirms previous reports which found no correlation of the folding energy to RNA half-life [159, 165]. RNA sta-

bility was also not correlated with gene length ($|r| < 0.09$; data not shown). Similarly, changes in RNA stability between growth conditions were not correlated to RNA folding energy or gene length, suggesting different attributes determine stability; perhaps the number of internal cleavage sites, 5′ or 3′ untranslated regions, susceptibility to different RNases or regulation by sRNAs.

3.3.2 Differentially Expressed Genes

The regulation of gene expression may be a functional role for the selective stabilization of mRNA transcripts. To confirm this hypothesis, we needed to determine which genes were regulated. A comparison of gene expression for cultures growing exponentially in the three media was performed. Three methods [185–187], each employing different underlying assumptions about gene expression, were used to predict statistically differentially expressed genes (DEG). Because each method gave diverse results, we took the common set of DEG—hereafter referred to as the "consensus set"—to be a conservative estimate of the DEG (see Fig. S7). Our observed fold change in RNA expression between acetate, MeOH and TMA are in good agreement with the previous, but limited, qRT-PCR and microarray studies that have been published [21] with a Pearson correlation coefficient, r, of 0.82 ($p < 3 \times 10^{-9}$; Supplemental Fig. S6A). We also found that, while absolute protein count was very weakly correlated to RNA expression (r=0.12, p>0.1), fold change in protein levels [16,17,33,34,104,199] were highly correlated to change in RNA expression (r=0.63, $p < 2.2 \times 10^{-11}$; Supplemental Fig. S6B).

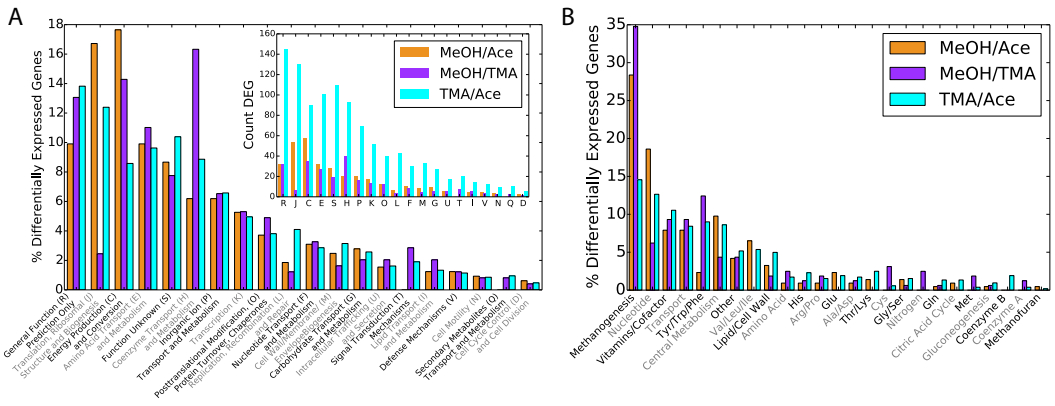


Figure 3.3: **Breakdown of Differentially Expressed Genes.** Breakdown of differentially expressed genes (DEG) comparing MeOH/Acetate (red), MeOH/TMA (green) and TMA/Acetate (blue) by (A) COG class and (B) metabolic subsystem (metabolic genes include those that are associated with reactions in the metabolic model *i*ST807). The inset in (A) shows the percent of DEG in each category and highlights coenzyme/vitamin metabolism (H) and translation and ribosome biogenesis (J) when comparing MeOH and TMA. Genes with a p-value ≤ 0.01 were considered to be differentially expressed.

Counts of differentially expressed genes can be found in Table 3.1. The consensus set consisted of 201±50 DEG comparing methanol and acetate (MA), 645±162 DEG comparing TMA and acetate (TA) and 211±62 DEG comparing methanol and TMA (MT) members with a p-value < 0.01 (see Fig. S7 for overlap between methods). The uncertainty in number of differentially expressed genes due to limited replicates was estimated to be 24-30% using a bootstrapping procedure (see Supplemental Section **Uncertainty in Differentially Expressed Genes** and Fig. S2). Genes involved in energy metabolism (C), translation (J), coenzyme metabolism (H), and amino acid metabolism (E) are most drastically effected, though all COG classes have at least one representative gene (Fig. 3.3A). We also computed differential

expression for gene operons. To do so, we combined reads over gene operons (as predicted by four different methods [112, 188–190]) and applied the same methods as for genes. We found that the consensus set DEG computed by operons largely reflected those computed by individual genes and that about 80% of genes were called as differentially expressed (see Table3.1).

Table 3.1: **Differentially Expressed Gene Statistics.** Count of genes that are differentially expressed when comparing growth substrates predicted by several methods.

Genes					
Comparison	edgeR	DESeq2	PoissonSeq	Consensus[b]	
MeOH vs Acetate	621	341	400	201 (126[d])	
MeOH vs TMA	2839	258	2348	211 (112[d])	
TMA vs Acetate	2762	757	1955	645 (335[d])	
Methylotrophic vs Acetotrophic	511	179	301	100	
Operons[a]					
Comparison	DOOR2	Microbes Online	ProOpDB	Rockhopper	Consensus[c]
MeOH vs Acetate	205 (144)	183 (110)	189 (129)	207 (151)	157
MeOH vs TMA	202 (148)	198 (117)	211 (143)	212 (158)	163
TMA vs Acetate	662 (450)	701 (343)	667 (406)	655 (469)	571
Methylotrophic vs Acetotrophic	112 (75)	97 (53)	91 (56)	112 (79)	76

[a]Intersection of the sets of differentially expressed genes predicted by the three methods. [b] Count of differentially expressed genes that are associated with reactions in the metabolic model. [c]The differential expression procedure applied to reads summed over putative operons of a specific dataset, where the number reported is total genes called as differentially expressed, while the number in parenthesis is the total number of operons called as differentially expressed. [d]Intersection of the sets of differentially expressed genes predicted to be in differentially expressed operons (because operon structures were not conserved across the method).

Differentially expressed genes were largely associated with metabolic reactions. Of the 201, 211 and 645 differentially expressed genes identified, about half were associated with reactions in our genome-scale metabolic models for *M. acetivorans* [31, 32] (affecting 149, 110, 359 total reactions, respectively). Genes associated with energy metabolism account for only 8-18%

of all DEG depending on compared substrates (Fig. 3.3A), demonstrating that carbon source plays a larger role in defining metabolic state than merely by changing expression of methanogenesis genes (Fig. 3.3B). Nucleotide, co-factor, aromatic amino-acid biosynthesis and transport metabolic pathways were also regulated extensively, each accounting for between 5 and 15% to total regulated genes (Fig. 3.3B). Notably, genes in translation also contribute to ∼3-16% of all differentially expressed genes, suggesting a tight coupling between growth substrate and genes affecting growth rate (Fig. 3.3A). In metabolism methanogenesis accounts for 15-35% of differentially expressed genes alone; however, genes involved in nucleotide, vitamin and cofactor biosynthesis as well as transport each constitute ∼10% of differentially expressed genes, suggesting that growth substrate has significant effect on the cell economy, likely affecting energy carrier balance and import and export of nutrients. Figure 3.4 shows a mapping of differentially expressed genes.

3.3.3 Regulatory Control Coefficients

In light of the fact that half-lives for mRNAs change significantly between growth substrates and that the changes are specific to certain functional classes of mRNA, it is likely that selective degradation of mRNAs plays a regulatory role. A recent theory was proposed to determine whether change in transcript abundance for a gene between two growth conditions is determined by change in the degradation or transcription rate [161,162]. Their theory defined "control coefficients" that describe the effective change in mRNA level as resulting from degradation or transcription, under the

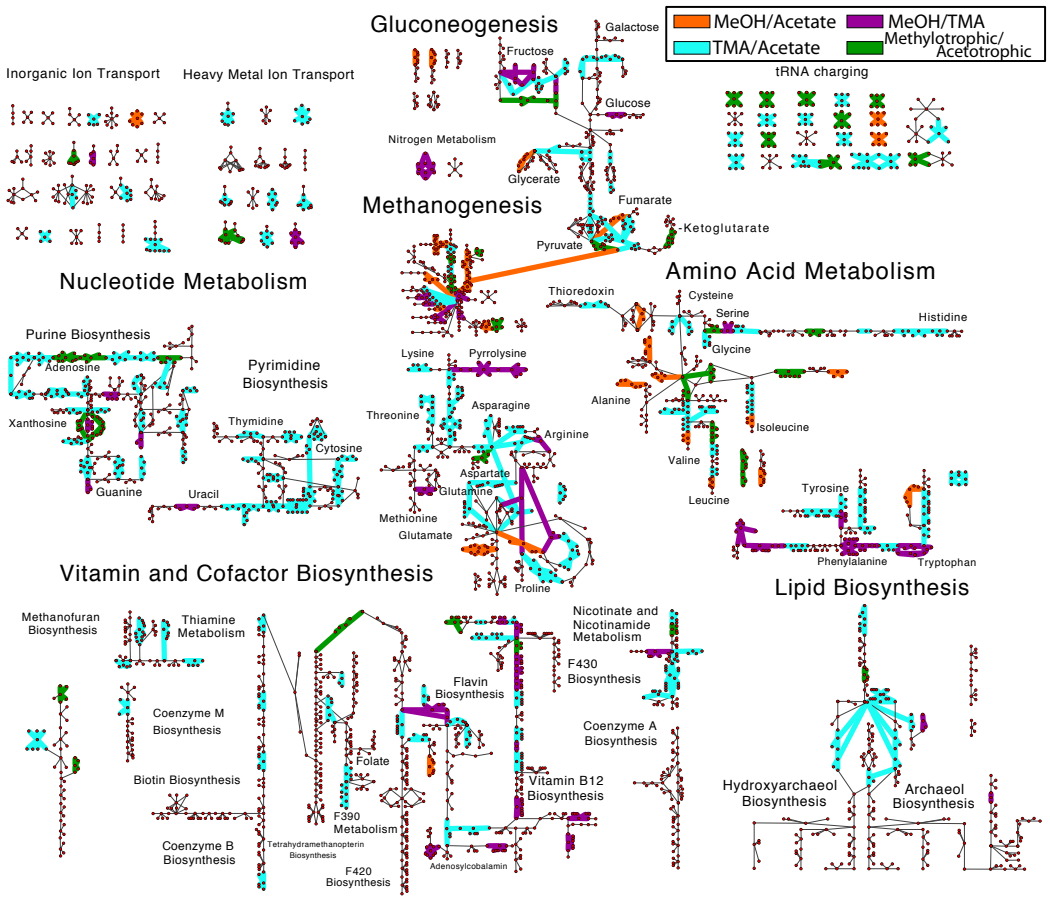


Figure 3.4: **Mapping of Differentially Expressed Genes (DEG) on Metabolism.** The map of all known metabolic reactions effected by differentially expressed genes comparing MeOH vs. acetate (orange), MeOH vs. TMA (purple), TMA vs. acetate (cyan) and MeOH/TMA vs. acetate (green). Reactions and metabolites are represented as green diamonds and red circles, respectively. Reactions are connected to participating metabolites by edges. Common metabolites are duplicated. Unregulated reactions are indicated by thin lines. Genes were considered differentially expressed if the p-value ≤ 0.01 as computed in all of the three methods: DESeq2, edgeR and PoissonSeq. Reaction and metabolite names can be seen by zooming into Figs. S19 & S20.

assumption that gene expression is at steady-state (i.e. the population is growing exponentially and in homogenous growth condition). We computed

the degradational ($\rho_D = -dln\gamma/dln[mRNA]$; where γ is the degradation rate) and transcriptional ($\rho_T = dlnk_{trn}/dln[mRNA]$; where k_{trn} is the transcription rate) control coefficients for each of the genes using the half-life and expression data. See Supplemental Section **Control Coefficients** for a derivation and description of these control coefficients. See Fig. 3.5 for a mapping of control coefficients comparing TMA and acetate.

Three regulation regimes are of interest [161–163]: 1) primarily degradationally controlled, ($\rho_D \geq 1$), 2) primarily transcriptionally controlled, ($\rho_D \leq 0$) and 3) mixed degradation and transcription control $0 < \rho_D < 1$. The results for all genes can be found in Tables 3.2. Our analyses show that between 16 and 28% of the changes in steady-state transcript levels are due to degradational control. Between 16 and 23% of transcripts show both degradational and transcriptional control effects. A summary of the control coefficients computed for differentially expressed genes can be seen in Table 3.3. Strikingly, more than 50% of all differentially expressed metabolic genes, defined as the 807 genes which are associated with reactions in the metabolic model (*i*ST807; see the next section), were primarily controlled by degradation. This high percentage underscores the substantial role that degradation plays in regulating metabolism in *M. acetivorans*.

3.3.4 Metabolic Model for *M. acetivorans*

Modifications to the metabolic reconstruction and model for *M. acetivorans* were necessary in order to simulate the effect of regulation on pathway usage. We updated the genome-scale metabolic reconstruction *i*MB745 [32] by

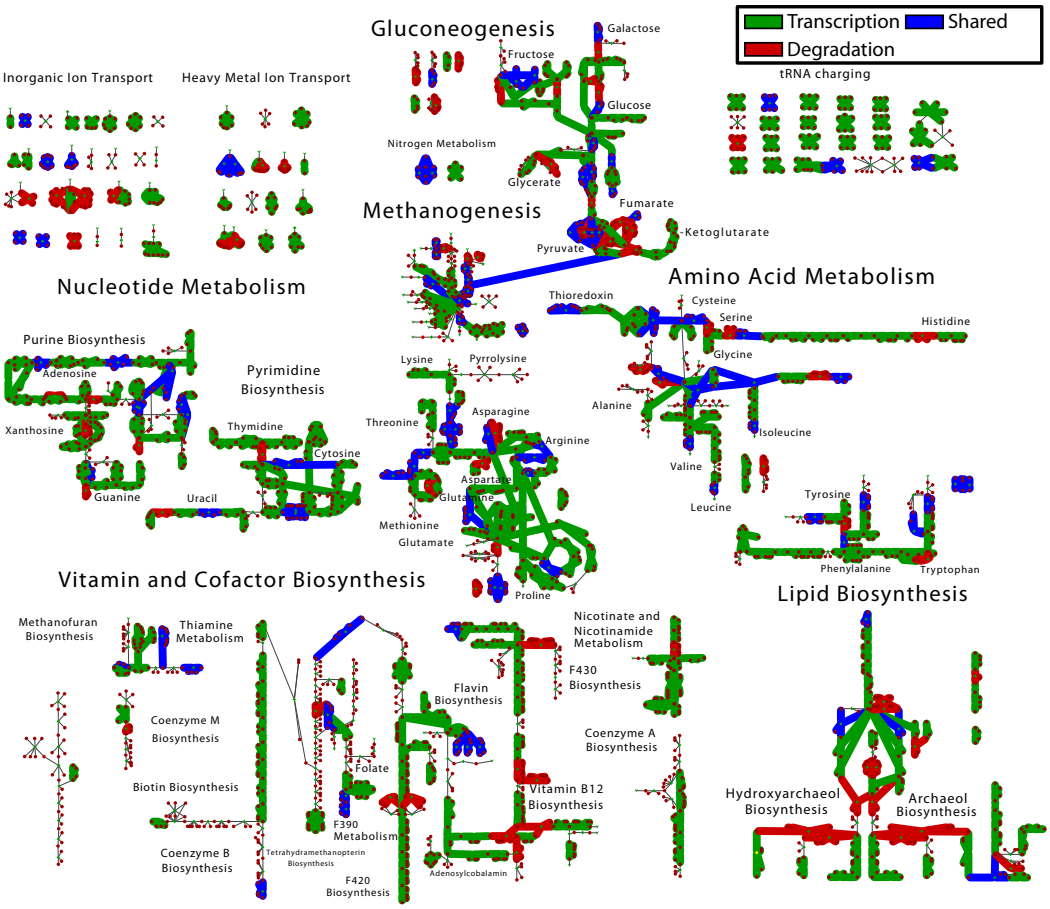


Figure 3.5: **Control Coefficients Mapped onto Metabolism.** A mapping of the control coefficients for changing mRNA expression levels between TMA and acetate. Red indicate reactions where mRNA levels are regulated by shifts in the degradation rate, while green indicates mRNA level shifts due to changes in transcription rate. Blue indicates reactions where mRNA levels are affected by both transcription and degradation rate. Reaction and metabolite names can be seen by zooming into Figs. S19 & S20.

incorporating newly characterized pathways and gene:reaction mappings. Several additional model improvements dealt with amino acid synthesis and ligation (e.g. tRNA-charging). First, the pathway for pyrrolysine biosynthesis was added to reflect the requirement for this amino acid for cells

Table 3.2: **Classification of Regulation Type; All Genes.**

	Transcriptionally Controlled $\rho_D \leq 0$	Shared Control $0 < \rho_D < 1$	Degradationally Controlled $\rho_D \geq 1$
Comparison	Number (%[a])	Number (%)	Number (%)
MeOH vs. TMA	2987 (67.2)	718 (16.1)	742 (16.7)
MeOH vs. Ace	2219 (61.2)	562 (15.5)	847 (23.3)
TMA vs. Ace	1763 (48.5)	844 (23.3)	1026 (28.2)

[a]Percentage of total genes for which half-lives and RNA reads were of sufficient quality to apply the control analysis (i.e. $CV_\tau < 0.5$).

growing on methylamine substrates (Fig. S8). Second, the alternate cysteine aminoacylation pathway, wherein *O*-phosphoserine is converted to cysteine while charged to $tRNA^{Cys}$, was added [92, 200] (Figs. S9 and S10). Third, tRNA-charging was explicitly included in the model, wherein the protein biomass composition was altered to require charged tRNAs instead of free amino acids. This change allows comparison of differential expression of tRNA genes with metabolic flux. Fourth, the biosynthesis pathway for methanofuran was updated based on recent characterization of the five *mfn* genes [201–204]. Methanofuran is a component in the methyl-oxidation branch of methanogenesis playing a role in a key redox step inter-converting a formyl group and CO_2 and is required for all modes of methanogenic growth [21, 24]. Fifth, metabolic pathways for incorporating the methylated sulfur compound, methylmercaptopropionate, have been added/updated based on recent molecular biology studies [25]. Finally, several genes and a recently characterized reaction have been added to lipid biosythesis [205–207]

The model biomass equation was refined by incorporating osmolytes, salts, and metals [208] and gluconeogenesis intermediates [209]. Additionally, pyrrolysine was added as a requirement when growth on methylamines,

Table 3.3: **Classification of Regulation Type; DEG.** Classification of regulation type for differentially expressed metabolic genes in the consensus set from Table 3.1. Metabolic genes are defined as those that are associated with a reaction in the metabolic model *i*ST807.

	Transcriptionally Controlled $\rho_D \leq 0$	Shared Control $0 < \rho_D < 1$	Degradationally Controlled $\rho_D \geq 1$	Indeterminate[b]
Comparison	Number (%[a])	Number (%[a])	Number (%[a])	Number (%[a])
MeOH vs. TMA	54 (25.6)	23 (10.9)	126 (59.7)	8 (3.8)
MeOH vs. Ace	25 (12.4)	49 (24.4)	97 (48.3)	30 (14.9)
TMA vs. Ace	107 (16.6)	149 (23.1)	320 (49.6)	69 (10.7)

[a]Percentage of DEG out of those for which half-lives and RNA reads were of sufficient quality to apply the control analysis (i.e. CV$<$0.5). [b]Regulation coefficients could not be determined because half-life data was low quality (CV$\geq$0.5).

allowing for the correct prediction of knockout *pyl* gene knockouts when growing on these substrates (Fig. S11). The model was modified to allow uptake of cysteine, a component of the media [178], at a rate consistent with experiments. Newly required osmolytes α-glutamate, *N*-acetyl-β-lysine and glycine betaine fix several dead-end pathways in the model, thus increasing the predictive capability of the model (see Discussion). Incorporating ion and metal requirements allows the model to predict how differential expression of membrane bound transporters affect osmolyte concentrations. Overall, the model can now take up most of the components of the Wolfe media [178] for which metabolites exist in the model (see Supplmemental Table S2).

Altogether the new metabolic reconstruction consists of 759 non-biomass reactions (829 when including metabolite exchange) with 807 associated genes. The reconstruction was laid out as a map to allow visualization of metabolic fluxes and gene expression data (see Fig. 3.4 and Fig. S12). The map is available in formats compatible with the Cytoscape [210] and Escher [211] software. The map and modified FBA model (called *i*ST807) are available in several formats in the additional files accompanying this manuscript. See Supplemental Section **Modifications to Metabolic Model** for a complete discussion of map and model modifications and verification.

3.4 Discussion

3.4.1 Regulation of Half-Lives

We found that *M. acetivorans* growing on different substrates exhibits drastically different RNA half-life stabilities. A much stronger growth rate effect was observed than in the previous studies; whereas half-lives in *E. coli* were shifted by a factor of 1.5 for a 6 fold change in growth rate, in *M. acetivorans* a nearly linear shift in half-lives with growth rate was observed. To test the hypothesis that half-life is correlated to growth rate, we scaled the half-life distributions by the doubling time, effectively defining a fraction of the cell cycle that an RNA persists before degradation (see Fig. 3.1 and Fig. S4). As demonstrated, the scaled half-life distributions align with means that are statistically indistinguishable ($p>0.33$, t-test). This scaling indicates that for a given growth substrate, the cell will modulate mRNA stability such that the half-lives are, on average, a constant fraction of the cell cycle. We could not identify any differentially expressed RNases among our data which would facilitate these changes in half-lives indicating that another mechanism is in play (perhaps sRNA, riboswitches, etc.).

Studies have examined how conserved mRNA half-lives are among related species. One study on two strains of *Bacillus cereus* showed high correlation among half-lives ($r=0.72$) [166] and another compared two species of the *Solfolobus* genus also finding high correlation ($r=0.51$) [168]. These studies show that RNA degradation is evolutionarily conserved and suggest

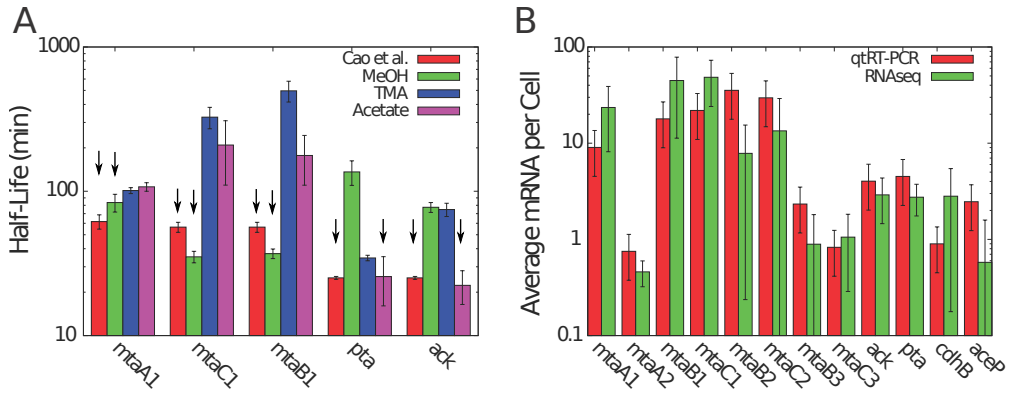


Figure 3.6: **Comparison of Transcripts with** *Methanosarcina mazei* **[212].** A) A comparison of mRNA half lives measured via our RNAseq data compared with a previous study using qtRT-PCR in the related organism *Methanosarcina mazei* growing in methanol or acetate. Cao et al. measured methyltransferase (*mtaA1*, *mtaCB1*) half-lives from methanol grown cells, while they measured acetoclastic gene (*pta*, *ack*) half-lives for acetate grown cells. As can be seen in the figure half lives match for methanol and acetate grown cells. Arrows indicate which bars correspond to the comparison to Cao et al. B) A comparison of mRNA copies per cell estimated via our RNAseq data, and previous studies that utilized RT-qPCR to quantify transcript abundance in the related organism *Methanosarcina mazei* grown in methanol (see Fig. S13 for acetate growth). Error bars are standard deviation of the mean for 3 replicates. Values from Cao et al. are for cells grown at 30°C compared to our cells which were grown at 37°C. All values agree within uncertanties except for *cdh*, *mtaA2*, and *mtaB2* indicating the organisms have similar expression profiles and our estimates for mRNA counts are good.

that our study of RNA stability in *M. acetivorans* may be extended to related organisms such as *M. mazei* or *M. barkeri*. We compared our measured half-lives to five that had been previously measured in *M. mazei* [212] and found them to have similar values (Fig. 3.6A). We estimated mRNA copies per cell for 12 transcripts in methanol and acetate growth conditions (Fig. 3.6B and Supplemental Fig. S13). Transcript counts also matched those measured in

M. mazei [212] suggesting these *Methanosarcina* species could have similar transcription and degradation characteristics, similarly to the two *Sulfolobus* species. In general some of the conclusions drawn from this study might hold for evolutionarily related methanogens (e.g. class II methanogens). However, half-lives of homologous genes are not correlated between distantly related organisms such as *E. coli* and *B. subtilis* [165], and therefore care should be used when extending the conclusions about individual transcripts here to distantly related methanogens (e.g. class I methanogens).

3.4.2 Inheritance of Gene Regulation

Our genome-wide study identified many DEG in addition to those previously identified due to the higher number of experimental replicates (higher statistical power of differential expression test) and the greater number of compared growth conditions. The current study verified 80% of previously [21] identified DEG in *M. acetivorans*, indicating that the RNA data is of sufficient quality to match potentially higher accuracy methods such as qtRT-PCR. Additionally, our pattern of DEG comparing methanol and TMA was similar to the one reported for *M. mazei* [213]. A total of 42 of the 71 directly homologous genes had similar patterns of differential expression in our dataset were highly correlation in fold change (r=0.85, $p< 10^{-5}$). The similarity of transcript abundance and half-life and similarity between differentially expressed genes (see Fig. 3.6) suggest that regulation is conserved among these closely related organisms. Genes that are similarly differentially expressed between these two *Methanosarcina* species include

methyltransferases for methanol and methylamine assimilation, a putative thiamine biosynthesis gene (*thiC*), genes involved in valine, leucine and aromatic amino acid biosynthesis (*aroDE*, *leuA*, and *trpABE*), cobalt metabolism enzymes (*MA1418* and *MA3250*) as well as many hypothetical proteins and regulators. The rest of the genes either had no homologs or were not determined to be differentially expressed. Genes that were not identified as differentially expressed could be optimized for the different environments in which *M. mazei* and *M. acetivorans* grow; perhaps due to the adaptation to freshwater and saline environments. A recent study of *M. mazei* strains along the Columbia River show differences in genomic content comparing those in fresh- and salt-water environments that resulted in differences in metabolic efficiency/utilization of TMA [214].

Genes coding for enzymes involved in biotin synthesis, including biotin synthase (*bioB*), were found to be very highly expressed ($>4\times$) when growing on TMA compared to the other substrates. This along with the observation that growth on TMA of the closely related methanogen *Methanohalophilus mahii* was stimulated by addition of biotin suggesting that it plays a role in methylamine-based growth [215], perhaps as a cofactor involved in vitamin and lipid biosynthesis.

3.4.3 Identification of Regulated Transcription Factors

Over 200 transcription factors have been putatively annotated in the DBD transcription factor database [216]. We found that 10, 9 and 13 of these transcription factors were significantly differentially expressed upon com-

paring MeOH vs acetate, MeOH vs TMA and TMA vs acetate, respectively. A number of the regulators have been characterized and are of particular interest. For example, the gene *msrA* (methanol-specific regulator A) was found to be highly expressed in both methylotrophic growth substrates confirming a previous report [17]. Additionally, *msrC* and *msrF* were found to be more highly expressed in TMA than acetate, also confirming previous experiments [17,18].

For the uncharacterized transcription factors we examined their expression characteristics to attempt to identify regulatory roles and targets (see Fig. S14 for correlations). Two putative transcription regulators (*MA2055* and *MA3302*) were more highly expressed in acetotrophic growth, the latter of which has previously been suggested to be a global regulator of methanogenesis pathways and dubbed *mreA* (*Methanosarcina* regulator of energy-converting metabolism A) [27]. The former of these is an uncharacterized MarR-like protein that has a similar gene expression profile to another transcription regulator *MA2212*, being correlated with genes involved involved in acetotrophic methanogenesis and ATP production (Fig. S14; Supplemental Section **Differentially Expressed Genes**). Two other *mre* like regulators were found to be differentially expressed: 1) *mreB (MA1671)*, when comparing TMA and acetate (and almost significant for TMA vs MeOH, p=0.0103), and 2) *mreD (MA3130)*, which was found to be more highly expressed in methylotrophic conditions and sits in a conserved cluster of genes that also contains *hdrABC*. The hdrABC homologs are differentially expressed under different growth conditions as previously reported [104]. These genes

reroute metabolic flux allowing them microbes to outcompete other organisms [104,217]. The proximity of *mreD* to *hdr* in the genome suggests it could regulate *hdrABC* differentiating methyltotrophic and acetotrophic growth, while *mreB* could differentiate methylamine growth from other conditions, potentially in optimizing one/several of the other *hdr* homologs; however, these hypotheses remain to be tested.

The putative nickel response regulator *MA1395* is highly conserved among all the methanogens and is anti-correlated to Hsp60 genes (*MA0086/MA1682/MA4413*) along with several key metabolic genes such as pyruvate synthase (*por*, *MA0031–MA0034*) and methenyltetrahydromethanopterin-cyclohydrolase (*mch*, *MA1710*), suggesting that during nickel starvation key metabolic enzymes are downregulated during methylotrophic conditions (see Fig. S14). It is, however correlated with quinolinate synthase genes (*MA0959/MA2716*), suggesting when nickel is taken up, more NADH/NADPH should be produced, and phoshoglycerate dehydrogenase (*MA0592*), which produces a precursor in the pathway that produces coenzymes F420 and F390. If *MA1395* does indeed regulate these genes it could act to sense available nickel in the environment and slow metabolism (via *mch* and *por*) while affecting redox balance (via production of NADH/NADPH and coenzyme F420). A previous study on regulation in *Methanococcus maripaludis* identified a homolog of *MA1395* (*MMP0719*) as being coexpressed with *mch* along with the methyltransferase *mtr*, the energy conserving hydrogenase *ehb* and genes involved in pyrophosphate uptake (*ppaC*) and purine biosynthesis (*purP*) [218]. Together these suggest an ancient role for *MA1395*

that needs to be further studied.

3.4.4 Regulation of General Transcription Factors

Our data is consistent with a previous study showing that the primary TATA binding protein (TBP; *tbp1*) transcript was similarly expressed under the three growth conditions [26]; however, it differs for comparisons of accessory TBPs wherein our data suggest that *tbp2* and *tbp3* are expressed at similar levels to each other, where the previous report showed that the latter of the two was much less highly expressed. Both [26] and our study show that *tbp3* is more highly expressed during methylotrophic than acetotrophic growth (almost identical fold changes in both studies), and this is supported by the observation that genes in amino acid metabolism and methylamine metabolism are differentially expressed upon its knockout as seen previously. Discussion of four additional transcription factors can be found in Supplementary Section **Differentially Expressed Genes**.

3.4.5 Regulation of Translation Machinery

During methylotrophic growth, proline, lysine and arginine tRNAs are more highly expressed as seen in the "tRNA charging" reactions in Fig. 3.4. Additionally, valine, alanine and methionine tRNAs are more highly expressed under methanol growth than acetate growth, and threonine more highly than during methylamine growth (see Fig. 3.4). Generally, they are 3–42$\times$ more highly expressed in methylotrophic conditions, suggesting either: 1) there is a much higher requirement for these amino acids under methylotrophic

growth, or 2) the slower growth in acetate can tolerate lower amounts of tRNAs, potentially due to the longer time allowed to find ribosomes while maintaining a protein production rate necessary for steady growth. Similarly, the genes coding ribosomal proteins are a factor of 8 times more highly expressed in methylotrophic growth conditions. These results lend additional support to the idea that cells differentially regulate ribosome numbers which have been shown in bacteria (*E. coli*; 6,800–72,000 depending on growth rate [219]) and archaea (*Haloferax volcanii*; 11,600–25,400 depending on growth rate [220]). This constitutes the first analysis of differential expression of translation machinery in *M. acetivorans*.

3.4.6 Regulation of Vitamin and Cofactor Metabolism

Vitamin and cofactor biosynthetic pathways include many differentially expressed genes (see Fig. 3.4), suggesting they play important roles in each of the growth conditions. Six genes involved in nicotinamide (*nadA1*), coenzyme F420 biosynthesis (*cofH1*, *mptA*), and cobalamin biosynthesis (*cbiX*) are more highly expressed during methylotrophic growth. These enzymes are found at the beginning of a linear pathways or at the branch-point of two pathways, allowing their regulation to have a large impact on production of cofactors. Additionally, 13 genes are most highly expressed in methylamine metabolism, including many in the pathway forming adenosyl-cobyric acid (*cbiCFHL*) and the final step thereof (*cobQ*), and those in heme production (*hemC*), riboflavin biosynthesis (*ribH*) and anthranilate synthase (*trpGE*). The gene involved in riboflavin biosynthesis is at a branchpoint of the coenzyme

F420 biosynthesis and cobalamin biosynthesis, and thus has the potential to divert metabolic flux, in the case of growth on methylamines, towards production of cobalamin, consistent with the fact that cobalamin biosynthesis pathway transcripts are highly expressed. We hypothesize that larger amounts of adenosylcobalamin are required for growth in the methylamines with one possible explanation being that there are three different methyltransferase systems encoded by *mtmCBA*, *mtbCBA* and *mtmCBA* which process tri-, di- and monomethylamine, respectively, abstracting a single methyl group each. If enzymatic activity does not vary significantly between the three enzymes, and therefore a similar amount of each protein exists in a cell to maintain a certain metabolic flux, three times the equivalents of cobalamin would be needed compared with growth on methanol. The differential expression data indicates that enzymes involved in cobalamin synthesis are in fact 2.5-3.5x more highly expressed in trimethylamine growth than in methanol growth, supporting this hypothesis. Biochemical characterizations could test this hypothesis.

3.4.7 Transcription/Degradation Control of Gene Expression

Many DEG in the consensus set were represented in gene-protein-reaction relations as part of *i*ST807, suggesting that the regulation due to different growth substrates could have large effects on the distribution of metabolic fluxes. A composite showing reactions affected by differentially expressed genes demonstrates significant regulation throughout metabolism (Figure

3.4). Key control points in linear pathways stand out, and we observe that regulation is generally clustered around branches in pathways (for example at the branchpoint between flavin biosynthesis and coenzyme F420 biosynthesis and extensively on the pathways leading from pyruvate towards different amino acids). Within the set of DEG, two obvious classes arise: methylotrophic and acetotrophic growth (contributing to 75% and 10% of the total variance computed via PCA; Fig. S1) which are classes with which to identify differential pathway usage.

Because the total concentration of transcripts associated with a gene are affected by both transcriptional rate and degradational rate the question of which factor plays the largest role is of interest. Dressaire *et al* [162] recently proposed a method to determine whether the level of a transcript is primarily controlled by degradation, transcription, or both and applied it to *L. lactis* and found that degradation played a role in setting transcript level for maximally 12% of genes studied. The method was subsequently applied to *E. coli* by Esquerré *et al* [161] showing a similarly small effect. In the latter case, a role of degradational control was found to play an important role in glycolysis, pentose phosphate, Entner-Doudoroff pathways and the TCA cycle. Furthermore, they identified a role of degradational control in setting the levels of key degradational machinery transcripts including several RNases and Hfq. Both of these studies, however, used chemostat experiments for cultures growing in one grow substrate, and the question remains whether degradational role plays a larger role in optimizing an organism for different growth substrates. We applied the analysis to generate the transcriptional

($\rho_T = dlnk_{trn}/dln[mRNA]$) and degradational ($\rho_D = dln\gamma/dln[mRNA]$) control coefficients, which describe the relative change in mRNA due to relative changes in transcription rate k_{trn} and degradation rate γ (see Supplemental Section **Control Coefficients**). In contrast to the previous single substrate experiments in *L. lactis* and *E. coli*, our analyses show that between 16 and 28% of the changes in steady-state transcript levels are due to degradational control (Table 3.2). A close analysis of the data leads to the striking conclusion that degradational control appears primarily at branchpoints and is enriched in amino acid metabolism, lipid metabolism and vitamin metabolism (Fig. 3.5 and Fig. 3.7A,C,& E). This localization at pathway branchpoins could indicate an important uncharacterized role for degradational control. And because more than half of differentially expressed metabolic genes appear to be controlled by change in degradation rate it is likely that the change in degradation plays a significant role regulating metabolism in *M. acetivorans* (see Table 3.3). If a regulated gene is significantly destabilized (stabilized), the production of its protein is expected to proportionally decrease (increase). This is because the average protein count for a gene should go as $\langle P \rangle \propto k_t k_r \tau / k_{dil}$ where k_t is the transcription rate, k_r is the translation rate, τ is the half-life and k_{dil} is the doubling rate. This argument neglects translational regulation with growth condition, which has been shown to exist in Eukaryotes [221] and in haloarchaea up to 30% of all transcripts [222, 223]. Translational regulation will additionally effect the k_r in this equation, potentially causing a nonlinear response when coupled with the change in k_t or k_d.

3.4.8 Modeling Metabolic Phenotype

The question then arises: How does the regulation of metabolic genes affect metabolism and what role does degradation play in defining metabolic state? To connect the regulation to metabolic function, we integrated the differential expression data with the updated genome-scale metabolic model to predict change in fluxes through metabolic pathways when the organism grows on different substrates. Briefly, the coefficients of biomass components, which describe a cell's physiological requirement for that molecule, were allowed to vary between growth substrates and fitted such that the deviation of the metabolic flux ratio from gene expression ratios of DEG in those pathways were minimized (see Supplemental Section **Additional Modeling Methods and Results** for a full description of the model; fitted biomass coefficients can be seen in Fig. 3.8 and Supplemental Fig. S15). Prior to this procedure, only flux changes in methanogenesis were correlated to expression data. After fitting the biomass coefficients flux ratios were more highly correlated to expression ratios and many more pathways carried different flux (Fig. 3.9, Figs. 3.10 and S16). The results of this procedure are hypotheses about which pathways carry more or less flux when the organism is grown in one condition compared to another. For instance, when comparing MeOH and acetate growth we find that in general most pathways carry significantly more flux when growing MeOH (Fig. 3.9, blue lines). Those reactions that carry more flux under acetotrophic growth outside of methanogenesis are primarily involved with biosynthesis of several amino acids (Ile, Thr, Trp,

Asn, Cys) and interconversion of alcohols and aldehydes (Fig. 3.9. red lines). The results comparing MeOH and TMA are more varied, especially with regards to ion and metal transport and carbohydrate metabolism (see Fig. 3.7F). By examining these results we can ascribe the function each differentially expressed gene has in defining the metabolic phenotype. The results contrasting the three substrates are mapped onto the metabolic network in Fig. 3.7B,D,&F. The largest effect is seen in acetate growth, where the majority of biosynthetic pathways carry less flux, especially those that generate amino acids and nucleotides as well as the cobalamine and coenzyme B pathways. Vitamin and cofactor metabolism show the majority of change compared to methanol. Decreased adenosylcobalamin and coenzyme F420 biosynthesis usage in acetate compared to methanol are associated with many differentially expressed genes (see previous section). Similarly, increased coenzyme F420 biosynthesis when growing on TMA compared to MeOH is associated with differentially expressed genes (*cbiCFHL*, *cobQ*). Our procedure correctly predicted the increase of coenzyme M in acetate grown cells compared with MeOH grown cells [224]; however, there are few studies that have examined biomass composition, and the validity of our predictions remains to be tested, especially from a quantitative perspective. The modeling results show how the DEG in this study effect metabolism connecting, for the first time, regulation to metabolic phenotype for this organism. A detailed analysis of flux changes and fitted biomass coefficients can be found in Supplemental Section **Additional Modeling Methods and Results**.

3.4.9 Conservation of Differentially Expressed Genes

The similarity of mRNA degradation rates among related species grown under similar conditions—as evidenced by comparison of our results to the limited studies on *M. mazei* [212] and findings of other studies in other organisms [166, 168]—suggests that metabolic control points and regulatory modes identified in one organism could be inferred in the metabolism of similar organisms. From an evolutionary stand-point related organisms subjected to similar environments would likely retain regulatory controls that optimize their fitness. A simple analysis shows that most of the differentially expressed genes are conserved among the *Methanosarcinae* (as shown in Fig. 3.11, and Supplemental Figs. S17 and S18). The amount of conservation drops off as one moves further away from *Methanosarcinae* and towards the simpler methanogens which lack significant metabolic capabilities that are found in the *Methanosarcinae*. This is clearly illustrated by the energy production (e.g. methanogenesis, electron transport) genes where the hydrogenotrophic methanogens lack significant portions of genes that are responsible for enabling the utilization of growth substrates beyond carbon dioxide. Despite the metabolic differences, a large fraction of all differentially expressed genes are still conserved across all the methanogens, especially those for translation. The problem that remains is to discover the structure and elements of the regulatory network and explore how they evolved.

3.5 Conclusions

In this study we have demonstrated that the carbon source regulates significantly more than just methanogenesis in *Methanosarcina acetivorans*, with genes all across the genome affected especially those involved in growth (e.g. transcription, translation) and metabolism (amino acid, nucleotide and vitamin biosynthesis). We found that while mRNA half-lives scale with doubling time, the effect was not uniform across functional classes, suggesting that the cell prioritizes certain capabilities at lower growth rates most likely to account for lower available energy. By combining the expression data with the half-life data we were able to identify genes that were likely targets of transcriptional regulation (e.g. transcription factors) or degradational regulation (e.g. sRNAs, RNases, riboswitches), providing testable hypotheses that can direct molecular studies of regulation within the organism. For example, we identified $\sim$32 putative regulators and their targets in *Methanosarcina acetivorans* and found that about 6 of the transcription factors and their targets were highly conserved across the order of *Methanosarcinales* and some were conserved across among all the methanogens. We hypothesized functions for those regulators based on the similarly conserved genes which can be readily tested with molecular biology experiments.

We found that most differentially expressed genes were involved in metabolism and about half of them were under degradation control. This is the first study to find such a prominent effect of degradation control in metabolism. As the RNases are not differentially expressed, this suggests

that *M. acetivorans* may make extensive use of sRNAs and riboswitches to fine-tune the degradation of mRNAs that encode metabolic proteins according to their environment. We tuned coefficients of the biomass reaction of our metabolic model to increase correlation between fluxes and expression data for each carbon source. By doing so we could determine how the metabolite demand along certain pathways varies with growth substrate. This procedure allows us to account for the differential expression of genes in those pathways. Altogether the half-life, differential expression, and correlated fluxes data allows us to build a richer picture of regulation than possible with transcriptome-only studies.

This work reveals many new features about regulation and metabolism in methanogens that inspired several hypotheses for further testing. Molecular genetic studies with corresponding transcriptomic information will be necessary to clarify the role of the differentially expressed transcription factors. Future bioinformatic and genetic studies will be required to confirm the presence and define the function of post-transcriptional regulators, especially any sRNAs. Additional proteomics data could confirm the changes in pathway fluxes that we predict for growth under various substrates. Such studies will allow the construction of a combined regulatory/metabolic network model that can predict how methanogens impact everything from the gut microbiome to the global carbon cycle.

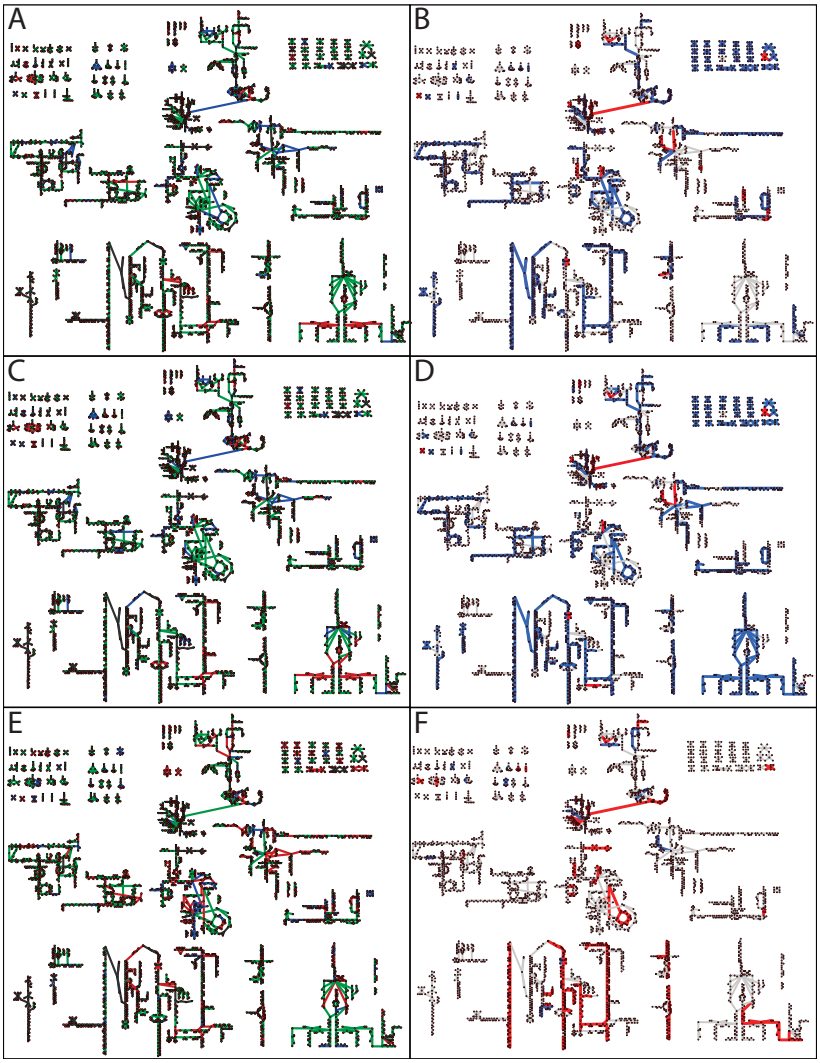


Figure 3.7: **Control Coefficients and Fluxes Contrasting All Substrates.** Comparisons of control coefficients to predicted metabolic fluxes (B,D,F). (A, B) MeOH vs Acetate, (C, D) TMA vs Acetate, (E,F) MeOH vs TMA. Control coefficients (A,C,E) indicate that reactions are transcriptionally controlled (green), degradationally controlled (red) or shared control (blue). Mappings of predicted metabolic fluxes indicate higher flux in the second substrate (red) versus lower flux in the second substrate (blue). Larger versions of A&B can be seen in Figs. 3.5 and 3.9. Larger versions of B-F can be seen in Figs. S19, S20, S21, and S22. The names for each reaction and metabolite can be seen by zooming into the the larger versions of the maps in the Supplementary Information.

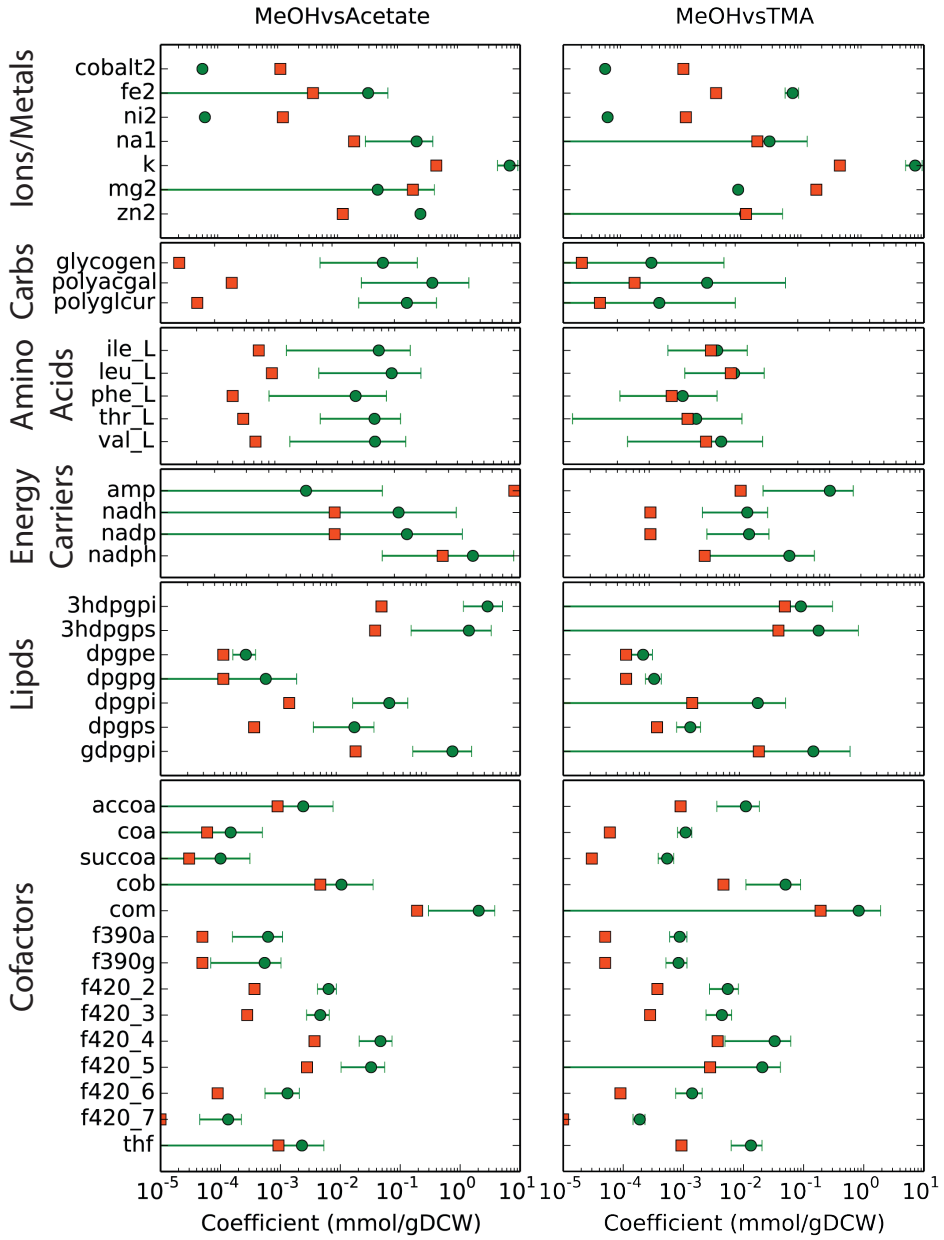


Figure 3.8: **Fitted Biomass Coefficients.** A comparison of fitted biomass coefficients. Orange squares indicate the coefficients for growth in MeOH while the green circles indicate the optimized biomass coefficients. Large error bars indicate that the coefficient can take on many values while still being optimal. Only metabolites with a significant shift comparing either MeOH to acetate or MeOH to TMA are included in the plot (all fitted biomass coefficients can be found in Supplemental Figure S15).

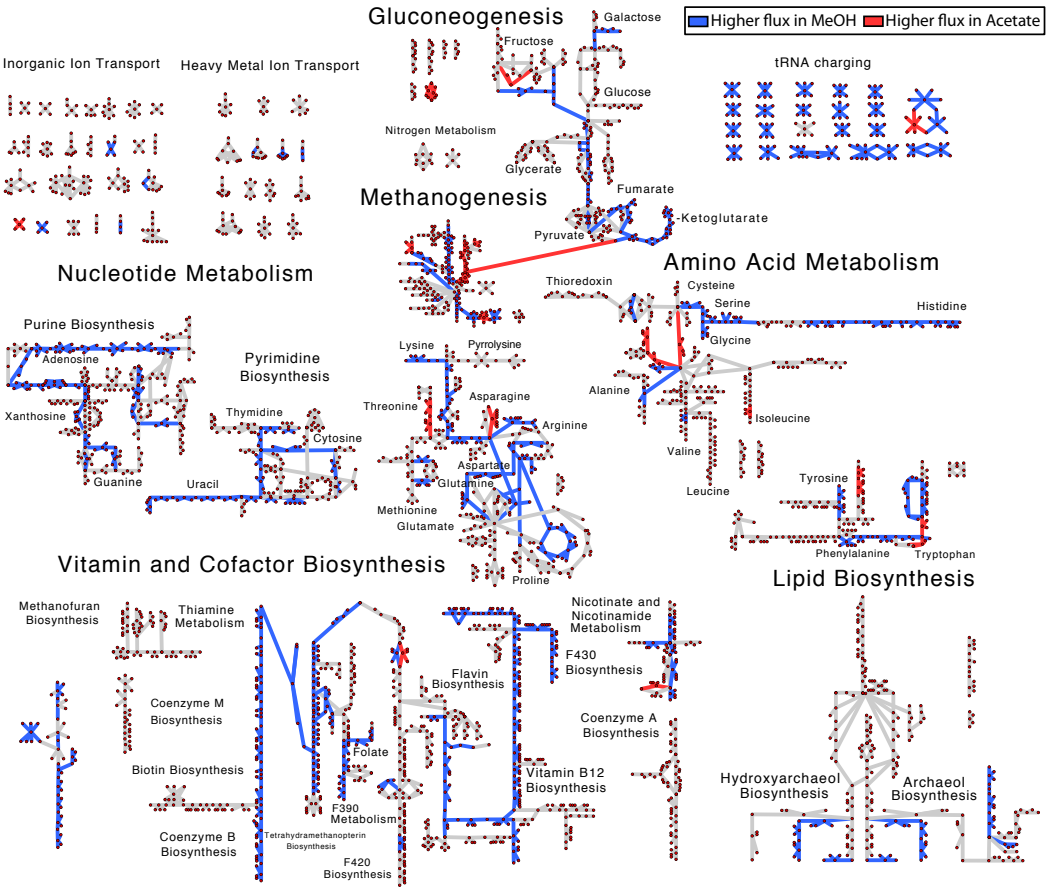


Figure 3.9: **Metabolic Flux Differences between MeOH and Acetate.** Predicted changes in flux comparing growth on methanol to growth on acetate. Pathways that carry more flux (>2 fold change in flux) when grown on acetate are indicated by red while those that carry more flux when grown on methanol are indicated by blue. Unaffected pathways are shown as grey lines. Reaction and metabolite names can be seen by zooming into Figs. S19 & S20.

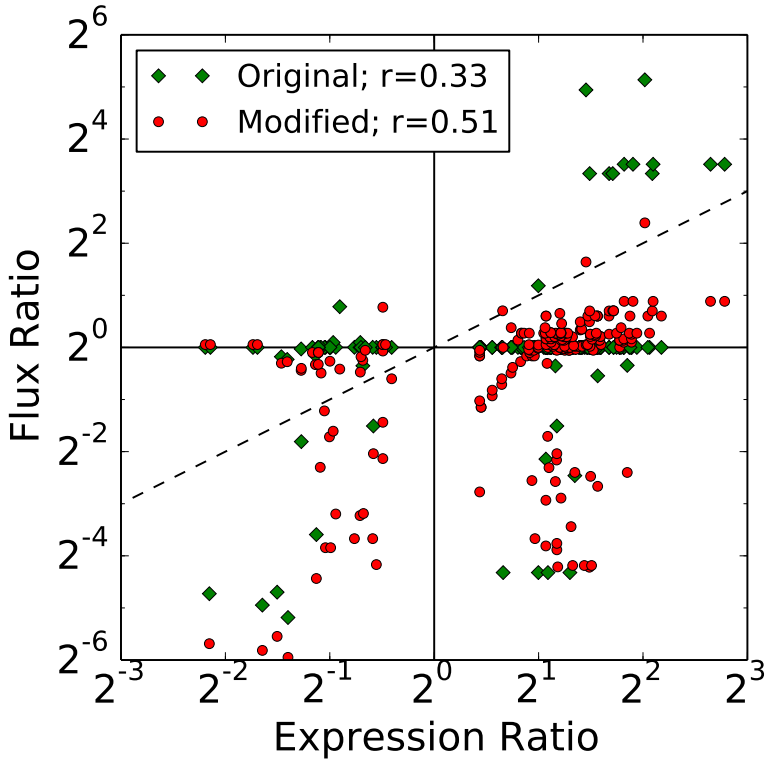


Figure 3.10: **Metabolic Flux vs Gene Expression.** Ratio of metabolic fluxes compared to ratio of gene expression for growth in different media. Each point represents a mapping between one reaction and one gene; therefore each reaction or gene may be represented by multiple points. If the same biomass requirements are used for the different growth substrates few of the reactions show any difference in flux (green diamonds) and there is weak correlation between expression and flux. The differences in fluxes that are observed are primarily due to genes encoding proteins that act in methanogenesis. By relaxing the assumption that biomass coefficient are constants across all growth substrates the model can be fit to improve the correlation between regulation and metabolism (red circles). After fitting many additional pathways are predicted to carry flux as demonstrated by more points moving off of the horizontal closer to $y = x$ (dashed line).

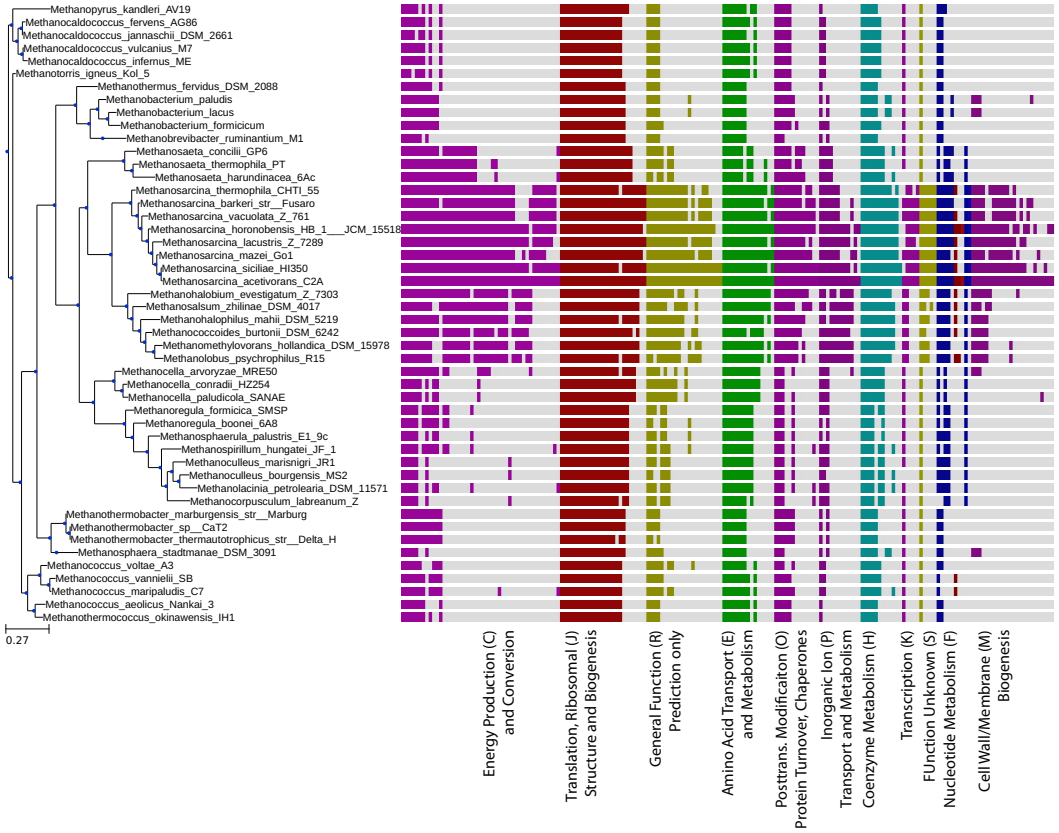


Figure 3.11: **Phylogeny of Differentially Expressed Genes.** Conservation of the genes that are differentially expressed between MeOH and acetate growth. Each vertical bar indicates that a homolog for the differentially expressed gene exists in the indicated species (computed as the bidirectional best hits functionality in the ITEP software [191] with an E-value cut-off of 10^{-5} for a database of $\sim$125000 proteins). Most differentially expressed genes are highly conserved among the *Methanosarcinales*; however a core set of genes are conserved across all methanogens.

3.6 Supplementary Information

3.6.1 Additional Methods and Materials

Each RNA sample came from an independent culture grown to exponential phase with an OD_{600} about 0.2 or 0.4 for slow (acetate) and fast-growing (MeOH, TMA) substrates, respectively. RNAseq was performed on a total of 69 RNA samples. Dataset descriptions and accession numbers may be found in Table S3.4. With the exception of two TMA samples, the library preparation was carried out using the ScriptSeq™v2 kit from EpiCenter. The two TMA samples were prepared with the Illumina TruSeq™v2 kit. Three samples for growth on methanol were generated in a previous report [25]. RNA isolation was performed as described in the methods section of the manuscript. Mapped reads in each sample were well fit by log-normal with a standard deviation of $\sim$1.0.

Table 3.4: **RNAseq Datasets.** A listing of all RNA-seq datasets used in the study. The first column designates the name of the sample file, followed by the growth condition, and the GEO database accession number for the sample.

Sample	Condition	Accession Number
	Steady-State[a]	
LK1_ATCACG_L007_R1_001	MeOH[c]	GSM2058125
LK9_TTAGGC_L003_R1_001	MeOH[c]	GSM2058137
LK17_GGCTAC_L004_R1_001	MeOH[c]	GSM2058150
PK19_CAGATC_L00M_R1_001	MeOH	GSM1569045

Table 3.4 (cont.)

Sample	Condition	Accession Number
PK20_ACTTGA_L00M_R1_001	MeOH	GSM1569046
PK21_GATCAG_L00M_R1_001	MeOH	GSM1569047
Metcalf_C2AM1_R1.PF	MeOH	GSM2058211
Metcalf_C2AM3_R1.PF	MeOH	GSM2058212
LK25_ATCACG_L003_R1_001	TMA[c]	GSM2058193
LK31_CAGATC_L004_R1_001	TMA[c]	GSM2058199
LK37_ATCACG_L005_R1_001	TMA[c]	GSM2058205
Metcalf2_C2AT_R1.PF	TMA[d]	GSM2058213
Metcalf2_C2AT_R2.PF	TMA[d]	GSM2058214
LK43_CAGATC_L006_R1_001	Acetate[c]	GSM2058164
LK49_ATCACG_L007_R1_001	Acetate[c]	GSM2058175
LK55_CAGATC_L008_R1_001	Acetate[c]	GSM2058187
RNA Degradation Study[b]		
LK2_CGATGT_L007_R1_001		GSM2058127
LK10_TGACCA_L003_R1_001	MeOH 5min	GSM2058139
LK18_CTTGTA_L004_R1_001		GSM2058152
LK7_ATCACG_L003_R1_001		GSM2058128
LK11_ACAGTG_L003_R1_001	MeOH 10min	GSM2058141
LK19_ATCACG_L005_R1_001		GSM2058154
LK3_TTAGGC_L007_R1_001		GSM2058130
LK12_GCCAAT_L003_R1_001	MeOH 20min	GSM2058143
LK20_CGATGT_L005_R1_001		GSM2058156

Table 3.4 (cont.)

Sample	Condition	Accession Number
LK4_TGACCA_L007_R1_001		GSM2058132
LK13_CAGATC_L004_R1_001	MeOH 30min	GSM2058145
LK21_TTAGGC_L005_R1_001		GSM2058158
LK5_ACAGTG_L007_R1_001		GSM2058134
LK14_ACTTGA_L004_R1_001	MeOH 60min	GSM2058147
LK22_TGACCA_L005_R1_001		GSM2058160
LK6_GCCAAT_L007_R1_001		GSM2058136
LK15_GATCAG_L004_R1_001	MeOH 120min	GSM2058148
LK23_ACAGTG_L005_R1_001		GSM2058162
LK26_CGATGT_L003_R1_001		GSM2058194
LK32_ACTTGA_L004_R1_001	TMA 20min	GSM2058200
LK39_TTAGGC_L005_R1_001		GSM2058206
LK27_TTAGGC_L003_R1_001		GSM2058195
LK33_GATCAG_L004_R1_001	TMA 30min	GSM2058201
LK40_TGACCA_L005_R1_001		GSM2058207
LK28_TGACCA_L003_R1_001		GSM2058196
LK34_TAGCTT_L004_R1_001	TMA 60min	GSM2058202
LK41_ACAGTG_L005_R1_001		GSM2058208
LK29_ACAGTG_L003_R1_001		GSM2058197
LK35_GGCTAC_L004_R1_001	TMA 120min	GSM2058203
LK42_GCCAAT_L005_R1_001		GSM2058209
LK30_GCCAAT_L003_R1_001		GSM2058198
	TMA 240min	

Table 3.4 (cont.)

Sample	Condition	Accession Number
LK36_CTTGTA_L004_R1_001		GSM2058204
LK75_TTAGGC_L005_R1_001		GSM2058210
LK44_ACTTGA_L006_R1_001		GSM2058166
LK50_CGATGT_L007_R1_001	Acetate 20min	GSM2058177
LK56_ACTTGA_L008_R1_001		GSM2058188
LK45_GATCAG_L006_R1_001		GSM2058168
LK51_TTAGGC_L007_R1_001	Acetate 30min	GSM2058179
LK57_GATCAG_L008_R1_001		GSM2058189
LK46_TAGCTT_L006_R1_001		GSM2058169
LK52_TGACCA_L007_R1_001	Acetate 60min	GSM2058181
LK58_TAGCTT_L008_R1_001		GSM2058190
LK47_GGCTAC_L006_R1_001		GSM2058171
LK53_ACAGTG_L007_R1_001	Acetate 120min	GSM2058183
LK59_GGCTAC_L008_R1_001		GSM2058191
LK48_CTTGTA_L006_R1_001		GSM2058173
LK54_GCCAAT_L007_R1_001	Acetate 240min	GSM2058185
LK60_CTTGTA_L008_R1_001		GSM2058192

[a]Datasets represent cells growing in exponential phase,

[b]Datasets taken at various times after transcriptional arrest,

[c]Datasets taken from the lifetime experiments prior to RNA degradation,

[d]These datasets were generated with the TruSeq v2.

Differential Expression Calling Procedures

A total of 16 datasets were considered: 8 replicates for MeOH, 5 for TMA and 3 for acetate. The zero timepoint from the degradation study accounts for three datasets in each growth condition. Additionally, 5 MeOH and 2 TMA replicates were obtained for exponentially growing cultures. Three methods were used for calling differentially expressed genes: DESeq2, edgeR and PoissonSeq. These three methods were chosen due to their relatively different assumptions about underlying distributions and method for normalization. Trimmed and mapped reads were loaded as datasets into each program without normalization. Multifactor statistical modeling was used when performing differential expression calling, where available, taking the growth substrate as the first factor and the library preparation method as the second factor. This step is necessary as 67% of the total variation observed in the experiment when multifactor design is not considered is due to the library preparation kit. When using the two factor design, the first principle component accounted for variation between methylotrophic and acetotrophic growth, and the second accounted for differences between methylotrophic growth substrates TMA and MeOH (Fig. 3.12).

Computations were performed using the R packages edgeR v3.8.5 [185], PoissonSeq v1.1.2 [186], and DESeq2 v1.6.3 [187]. Genes were considered differentially expressed when the p-value was < 0.01. All differentially expressed genes can be found in Supplementary Table "DifferentiallyExpressedGenes.MultiFactor.xlsx". All R scripts used to perform differential expression calling may be found in the Supplementary File "DEGComputa-

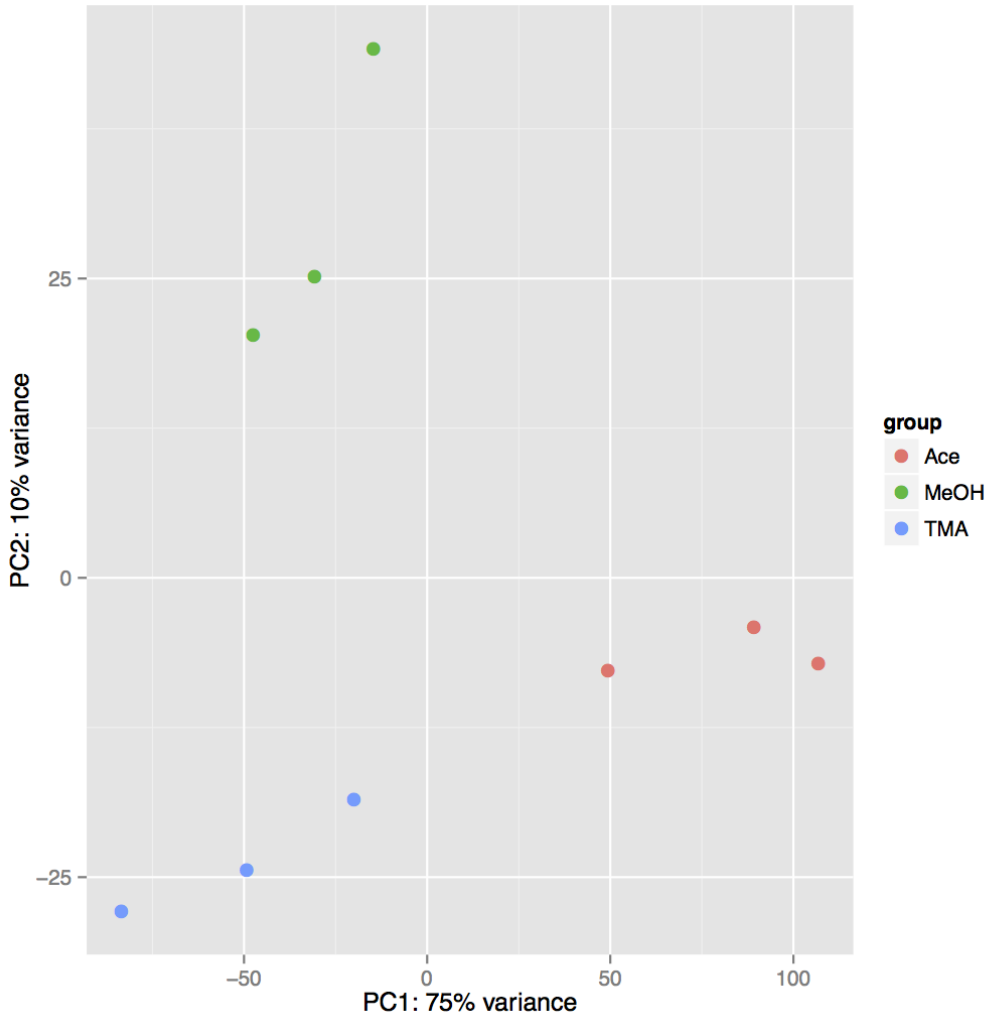


Figure 3.12: **PCA for RNA-Seq Datasets.** Principle component analysis results of the zero time RNA expression datasets computed via DESeq2. The first component separates methylotrophic and acetotrophic growth accounting for most of the variation seen, while the second component distinguishes the two methylotrophic substrates. The PCA reaffirms the traditional classification of acetotrophic and methylotrophic growth as being orthogonal.

tion.zip". Differential expression calling procedures were as follows:

DESeq2 – The library preparation was defined as the first experimental

factor, and growth substrate as the second factor. DESeq2 was run using the parallel implementation and the option "addMLE=TRUE", contrasting the growth conditions. Differential expression statistics were sorted by the adjusted p-value (Benjamini–Hochberg method) and stored to file. Additionally, a variance stabilizing transformation was performed on the data before a PCA analysis. The first two principle components were plotted to show the separation by growth types (Fig. 3.12).

edgeR – The library preparation was defined as the first experimental factor followed by the growth condition. Normalization factors were computed and a generalized linear model (GLM) was estimated. The dispersion trend over multiple genes was then calculated followed by the per gene (tag–wise) dispersion. Finally, the GLM model was fit, and the adjusted p-value (Benjamini–Hochberg method) was computed and the data stored.

PoissonSeq – Total mapped reads for each gene were scaled by a factor 0.1 so that the PoissonSeq method did not overflow. The differential expression calling routine was run with the "pair parameter" set to false and the data type taken to be "two–class" (substrate and library preparation kit). A total of 100,000 permutations were performed per comparison. Data was sorted by adjusted p-value (using the PoissonSeq default method: permutation plug–in) and stored to file.

Uncertainty in Differentially Expressed Genes

A nonparametric bootstraping approach was used to estimate the uncertainty in the number of differentially expressed genes. Briefly, the DESeq2 work-

flow was applied to subsets of the all the RNAseq data sets and the counts of DEG were enumerated. All combinations of sets of RNAseq data ranging from two to the maximum number of replicates were generated for each growth condition (MeOH, TMA and Acetate). The Cartesian product of these sets were generated, and the DESeq2 workflow was used to estimate the total number of differentially expressed genes between the three conditions. For a given total count of datasets (N_{MeOH}+N_{TMA}+$N_{Acetate}$), the average and standard deviation in the number of genes called as differentially expressed with confidence p$\leq$0.01 were computed. Mathematically,

$$\forall c \in \{\text{MeOH}, \text{TMA}, \text{Acetate}\} \tag{3.4}$$

$$\forall i \in \{2, ..., N_c\} \tag{3.5}$$

$$S_c = \bigcup_i \left\{ \binom{N_c}{i} \right\} \tag{3.6}$$

$$P = S_{\text{MeOH}} \times S_{\text{TMA}} \times S_{\text{Acetate}} \tag{3.7}$$

where c is the condition, N_c is the count of datasets measured in that condition, S_c is the set of all combinations containing 2 to N_c datasets and P is the product of all unique datasets. This accounted for about 26,000 sets of differential expression combinations. At each $N \in 2, N_c$ the coefficient of variation in number of DEG was computed (Fig. 3.13). This analysis demonstrates that the CV decreases as additional datasets are included. The CV is a measure of the uncertainty in the set of DEG; we estimate an uncertainty of 24-30% when using all 16 datasest.

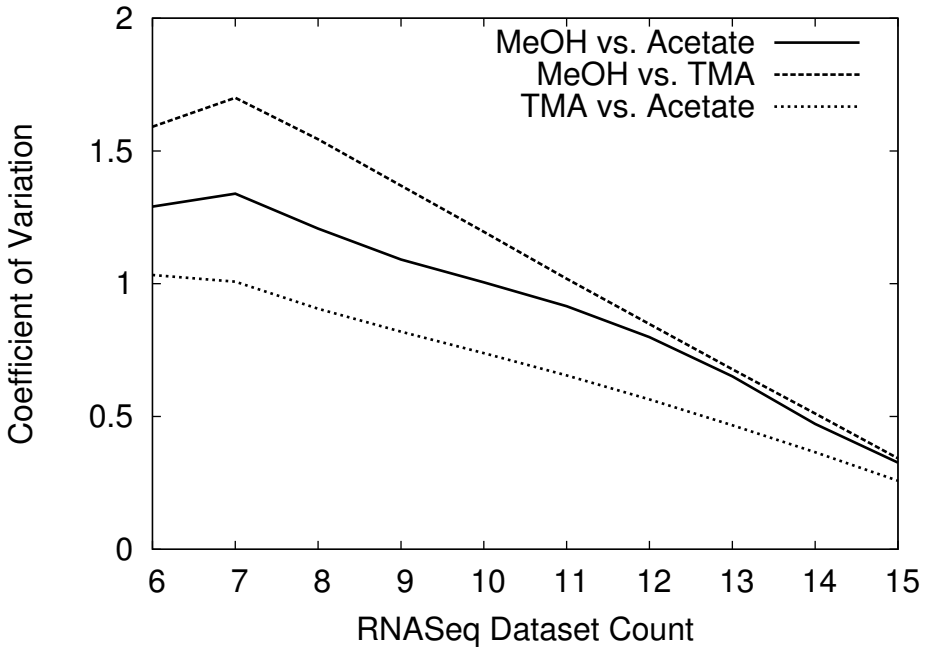


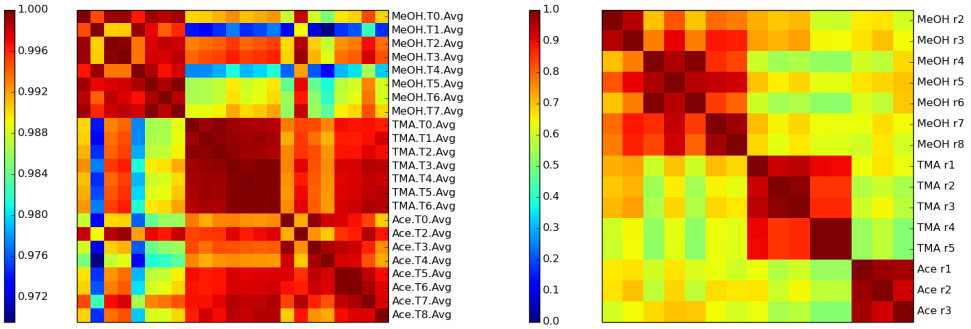
Figure 3.13: **Coefficient of Variation of Differentially Expressed Genes.** Estimated coefficient of variation in the number of differentially expressed genes as a function of the number of RNAseq datasets considered as computed in Section S3.6.1. The CV decreases as the number of datasets increases, suggesting the method become converge on a consistent set of differentially expressed genes. Using these curves we estimate a CV of about 0.25–0.3 when using all 16 datasets.

3.6.2 Additional Experimental Results

Half-Life Data

To assess reproducibility of the experimental procedure, correlations were computed across timepoints. The profiles were averaged across all three

replicates before correlations were computed. As can be seen from the correlation matrix in Fig. 3.14a, MeOH, TMA, and acetate correlate highly within condition and cluster together.



(a) (b)

Figure 3.14: **Pearson Correlation Matrices.** (A) Comparison at various time points after transcriptional arrest. Each pixel represents the correlation between the averages of from three replicate measurements, computed over the full expression profiles. (B) Comparison of data normalized using the procedure of DESeq2 [187] for the 8 methanol, 5 TMA and 3 acetate replicates. The three growth conditions form obvious clusters, supporting the idea that the experiments were reproducible. Color bars shows the magnitude of the correlation coefficient.

Limited data for half-lives of genes in *Methanosarcina* currently exist. Only half-lives for 5 genes have been reported from the related organism *Methanosarcina mazei* zm-15 [212] where they measured the half-lives of the methyltransferase genes (*mtaA1*, *mtaCB1*) from methanol grown cultures and acetoclastic genes (*pta*, *ack*) from acetate grown cultures. As can be seen in Figure 3.6A, the half-lives we measured agree quite well and within one order of magnitude in the worst case. These differences might be due to uncertainty in the measurements, differences inherent to the different organisms, and

the fact that Cao et al. measured half-lives at 30°C—and showed that there are different temperature stabilities for transcripts—while we measured half-lives from cells growing at 37°C. In general, they agree quite well and provide confidence in our measurements. The comparison also highlights an interesting fact, that the different transcripts have different stabilities after being grown in different conditions. This supports our hypothesis that half-lives are differentially stabilized/destabilized in different conditions and wy the control coefficients are important to compute, likely by post-transcriptional modification as Cao et al. concluded [212] or via small RNA regulation.

A scaled value for each of the half-lives was computed for each condition. The scaled half-life value is computed as $HLs = HL/(DT \times 60)$ where the HL is the unscaled half-life and DT is the doubling time in that condition in units of hr. Growth rates used for scaling of data for were taken as the average of experimentally determined values reported in literature. For MeOH, TMA and acetate, growth rates used were 7.5hr [104, 134, 194], 8.9hr [104] and 24.6hr [104,194,195], respectively. This scaled half-life represents the fraction of the cell cycle that a transcript will remain stable. As Fig. 3.15 demonstrates, regardless of the growth substrate, on average the RNA molecules persist for about the same fraction of the cell cycle. The scaled half-lives are statistically the same across all conditions ($p>0.33$, t-test) with an average value of 12.7%$\pm$3.5% of the cell cycle. Because growth rate is linearly proportional to ATP production rate [32], and it is generally assumed that growth rate and growth yield are co-optimized in prokaryotic organisms [225], we can

hypothesize two scenarios that cause this constant fraction of the cell cycle. First, the cell is optimized to use as little ATP as possible while maintaining a level capable of allowing the translation of new proteins at the correct rate and thus is linearly proportional to the ATP production rate. Since RNA turnover is a trade-off between degradation rate and production rate, the latter of which should be directly proportional to the energy required to ligate nucleotides into new RNA molecules, their steady-state values should be proportional to ATP production. An alternative but related second scenario is also possible; namely, that RNA maintains a constant fraction of the total cellular weight regardless of the growth rate. In this scenario, it is not the ATP consumption requirement in RNA production, but instead the production and degradation kinetics that set the steady-state amount of RNA in a cell. Since cell mass is proportional to ATP production rate and growth rate, the steady-state RNA is indirectly related to ATP production rate through the maintenance of constant mass fraction. One might argue that because the ATP cost of RNA production is such a low fraction of total energy expenditure, the latter of these two explanations is more likely.

Differentially Expressed Genes

The coefficient of variation (CV) computed using our uncertainty estimation method (Supplemental Section **Uncertainty in Differentially Expressed Genes**) decreases as the number of RNAseq dataset replicates increase (Fig. 3.13). The falling CV is consistent with prior studies that showed for DESeq (the precursor to DESeq2) as well as other similar methods such as edgeR

and PoissonSeq, the sensitivity rate (fraction of true positives) increases with number of replicates [227, 228]. At 15 datasets, the CV is about 30% of the mean. Linearly extrapolating the trends to 16 datasets results in a CV of 24% of the mean. Therefore, we estimate that the uncertainty in the number of differentially expressed genes is between 24 and 30% of the total number.

Among the differentially expressed genes, four putative regulatory proteins stood out: *MA0866*, *MA1395*, *MA2212* and *MA4346*. We compared the expression profiles across the three substrates to DEG that had nearly the same pattern of conservation (± 2 genes) and these genes were found to be highly correlated/anticorrelated to the regulators(Fig. 3.25). Analysis of these similarly conserved genes leads to interesting predictions that the regulators could be either directly regulating the group of genes, or is coregulated with them by another transcription factor.

The first highly conserved regulator *MA0866* encodes a PhoU type protein that likely plays a role in phosphate uptake. As expected, its expression is highly correlated with a phosphate related genes including a phosphate transporter subunit (*pstS*, *MA0889*), nicotinate phosphoribosyltransferase (*pncB*, *MA2533*) as well as TCA cycle enzymes citrate synthase (*MA0249*) and malate dehydrogenase (*mdh*, *MA0819*). Additionally, it is anticorrelated to the gene responsible for the final step of lysine synthesis (*lysA*, *MA0762*). These results suggest the gene could play a role in maintaining phosphate and energy balance in the cell, if it were to regulate these enzymes.

MA2212 is a TrmB-like regulator. It is notable because it is correlated highly with acetotrophic genes *ack* (*MA3606*), *pta* (*MA3607*), and *cam* (*MA2536*)

as well as subunits of the ATP synthase (*MA2433/MA2435/MA2440*)

The final regulator with high correlation to genes with similar conservation *MA4346* was specific to the family *Methanosarcina* but most genes that were similarly had nonspecific or no annotated function.

Estimating mRNA Levels

An estimate of the average copy number of each mRNA in an average cell, N_i for each of the three growth substrates were computed using the following equation,

$$N_i = x_{\text{RNA}} \cdot \rho_{\text{cell}} \cdot V_{\text{cell}} \frac{a_i}{m_i} \tag{3.8}$$

subject to the constraint

$$m_{RNA} = \sum_{i}^{N} a_i m_i \tag{3.9}$$

where a_i is the fraction of total mRNA mass m_{RNA} that the transcripts from a single gene accounts for, which is taken to be linearly proportional to the RPKM values from the RNAseq data; ρ_{cell} is the density of an *E. coli* cell taken from the CyberCell Database [229]; V_{cell} is the volume of the cell computed from our previous characterization of cell dimensions [35]; x_{RNA} is the mass fraction of total cell mass that is RNA, which is taken from the metabolic model (24% of total cell dry mass) [32]; N is the total number of

genes considered in the analysis; and m_i is the molar mass of the transcript of interest. Since the volume of cells grown in TMA were not measured, but the growth rates are similar to those for cells grown in MeOH, the volume were assumed to be the same. Total RNA for a single "average" cell was estimated to be 23.8 and 10.6 fg from MeOH/TMA and acetate grown cells, respectively. The counts of each RNA estimated using this analysis for the three substrates can be found in Supplementary File "EstimatedRNACounts.xlsx".

As a quality check, the numbers computed through this analysis were compared to those that were reported in Cao et al [212] measured for *M. mazei* growing in MeOH and acetate using RT-qPCR as shown in Fig. 3.24. In Cao et al [212], the values were originally reported per 100,000 16s-rRNA transcripts. We rescaled these numbers to be proportional to 14,000 rRNA transcripts, the average number of ribosomal protein Rpl18p count we measured previously [35] using a single-molecule pulldown [103]. As can be seen in the figure, all transcript counts except *mtaA2*, *mtaB2*, and *cdhB* are statistically indistinguishable, suggesting that estimates for mRNA numbers are good. The minor disagreement for the three transcripts could be due to the difference in the two species.

Control Coefficients

The confounding effects of changes in transcription and degradation rates on average mRNA level that occur with changes in growth rate can be deconvoluted. We attempt to estimate the effect of each by using a recently reported method that computes the extent that transcription and degradation

have in setting the steady-state level of mRNA in a cell [161–163]. The analysis is based on the assumption that the cell is at steady-state, implying the transcription rate and the degradation rate are balanced, or

$$k_{trn} = \gamma \cdot M + \mu \cdot M, \tag{3.10}$$

where k_{trn} in the transcription rate, γ is the mRNA's degradation rate constant as computed from RNAseq data, μ is cell growth rate and M is the average copy number of a transcript. The transcription rate—which is a proxy for change in growth rate (ribosomal count, etc.) and changes in promotion or repression of a gene—and degradation rate—changes due to active or passive degradation by RNAses or post-transcriptional control by sRNA—are computed per mRNA. Writing down the total differential and manipulating, the contribution of each process can be estimated as

$$dM = \frac{dk_{trn}}{\gamma + \mu} - k_{trn}\frac{d\gamma}{(\gamma + \mu)^2} - k_{trn}\frac{d\mu}{(\gamma + \mu)^2}. \tag{3.11}$$

Rearranging and noticing that at steady state, $M = k_{trn}/\gamma$ yields

$$\frac{dM}{M} = \frac{dk_{trn}}{k_{trn}} - \frac{d\gamma}{\gamma + \mu} - \frac{d\mu}{\gamma + \mu} \tag{3.12}$$

which can then be written as a relation between relative changes in transcript count due to changes in each rate,

$$1 = \frac{dlnk_{trn}}{dlnM} - \frac{dln\left(\gamma + |\mu|\right)}{dlnM} - \frac{dln\left(\mu + |\gamma|\right)}{dlnM}, \tag{3.13}$$

where here we use $|\cdot|$ to denote that this value is held constant in this term. Two cases for this equation can be considered: 1) the degradation of mRNA due to dilution is negligible, $\gamma >> \mu$, and 2) degradation due to dilution cannot be neglected. In the former, the contributions are due to transcription ρ_T and degradation ρ_D, or

$$1 = \frac{dlnk_{trn}}{dlnM} - \frac{dln\gamma}{dlnM} \tag{3.14}$$

$$= \rho_T + \rho_D. \tag{3.15}$$

In the latter, the contributions are due to transcription ρ_T, degradation ρ_D and growth (dilution) ρ_G, or

$$1 = \frac{dlnk_{trn}}{dlnM} - \frac{dln(\gamma+|\mu|)}{dlnM} - \frac{dln(\mu+|\gamma|)}{dlnM} \tag{3.16}$$

$$= \rho_T + \rho_D + \rho_G. \tag{3.17}$$

The majority of mRNA satisfy $\gamma >> \mu$. Therefore, we proceed neglecting ρ_G. Dressaire et al. applied Eqn. 3.14 to *L. lactis* growing in chemostats at different rates and found that only a few percent of genes are degradationally controlled [162]. Esquerré et al. applied the same analysis to *E. coli* growing in chemostats at several different rates [161]. Both studies ignored the dilution effects, citing the small average half-lives relative to the doubling times studied (1-8%) similar to our average 12.7%. In contrast to these studies, we found a significantly higher number of genes that appear to be degradation-

ally controlled (16-28%). This percentage was even higher when considering only genes associated with metabolic reactions (48-60%). Control coefficients calculated between each of the three growth substrates can be found mapped onto the metabolic network in Figs. 3.30 and 3.31.

3.6.3 Additional Modeling Methods and Results

Modifications to metabolic model

COBRAPy [230] was used to handle the flux balance computations and all changes to the metabolic model. The *M. acetivorans* model (*i*MB745) [32] required additional improvements in order to accurately predict the metabolic behavior when grown in the standard high-salt medium [178] used for the RNA seq experiments. All components of this medium that could be taken up by the metabolic model are listed below and were turned on with a default lower bound of -1000 $\frac{mmol}{gDW \cdot h}$.

Table 3.5: **Methanogen Growth Media.** Components of high-salt medium used to grow methanogens [178]. Starred (*) metabolites currently do not have exchange uptake reactions in model.

Core components	NaCl, MgCl, $CaCl_2$, $NaHCO_3$, KCl, KH_2PO_4, NH_4Cl, Na_2S*, Cysteine, resazurin*
Trace elements	$Fe(NH_4)_2(SO_4)_2$, $CoCl_2$, $MnSO_4$, Na_2MoO_4, Na_2WO_4, $ZnSO_4$, $NiCl_2$, $CuSO_4$, H_3BO_3*, Na_2SeO_3*, nitrilotriacetic acid*
Vitamins	p-aminobenzoic acid, Ca pantothenate, riboflavin, thiamine HCl, biotin, folic acid, vitamin B_{12}, pyridoxine HCl*, α-lipoic acid*, nicotinic acid*

In addition to refining the methanofuran biosynthesis pathway, the alter-

nate aminoacylation pathway for cysteine and the pyrrolysine biosynthesis pathways—two evolutionarily significant pathways—were added to the model. Addtionally, reactions to allow uptake of methylmercaptopropionate (MMPA) were added, and the gene-reaction-protein rules for methylated sulfur compound metabolism were significantly revised according to new genetic evidence [25]. The new metabolic model *i*ST807 maintains its predictive capability from the previous model and reproduces the methanogen's inability to grow on methylamine substrates without pyrrolysine. The resulting model growth rates, methane production, and carbon dioxide production are in good agreement with the experiments (Fig. 3.22).

Alternate cysteine aminoacylation pathway In 2005, O'Donoghue, et al. [92] predicted the existence of the alternate cysteine aminoacylation pathway within a handful of methanogens, including *M. acetivorans*, which was later confirmed [200]. This indirect charging pathway for cysteine, shown in Fig 3.19, is unique to archaeal species. In certain methanogens, such as *M. jannaschii*, it is the only mechanism to charge cysteine onto its tRNA. Many archaeal species obtained the canonical cysteine charging pathway through horizontal gene transfer, explaining the presence of both pathways within *M. acetivorans*. Since *i*MB745 did not include this evolutionarily significant alternate cysteine charging pathway, it was incorporated into the modified metabolic model.

Briefly, SepRS acylates the precursor O-phosphoserine onto the cysteinyl-tRNA, then SepCysS converts the acylated O-phosphoserine into cysteine.

Table 3.6: ***Methanosarcina acetivorans*** **Biomass.**

Reactant	Value	Units	Subcomponents
ATP	44.1	mmol/gDCW	
H_2O	44.1	mmol/gDCW	
Protein	0.63		0.85Leu + 0.65Gly + 0.59Lys + 0.21Met + 0.4Asn + 0.36Pro + 0.4Phe + 0.62Ser + 0.0808Pyl + 0.49Thr + 0.15His + 0.4Arg + 0.33Tyr + 0.48Asp + 0.62Val + 0.11Cys + 0.09Trp + 0.23Gln + 0.71Glu + 0.62Ala + 0.66Ile
RNA	0.24		0.58dATP + 0.44dCTP + 0.44dGTP + 0.59dTTP
Lipid	0.05		0.005DPGPG + 0.2143HDPGP + 0.027DPGPI + 0.2873HDPGPI + 0.005DPGPE + 0.0573HDPGPE + 0.011DPGPS + 0.2443HDPGPS + 0.148GDPGPI
DNA	0.04		0.48ATP + 0.42CTP + 0.5GTP + 0.6UTP
Trace Metabolites	0.04		0.24Putresine +0.044Homospermidine +0.0009AcCoA +0.0000CoA +0.0204NAD^+ +0.00093NADH +0.00093$NADP^+$ +0.00371NADPG +0.00003SucCoA +0.00929AMP +0.191Coenzyme M +0.00037Coenzyme F420-2 +0.00028Coenzyme F420-3 +0.00371Coenzyme F420-4 +0.00279Coenzyme F420-5 +0.00009Coenzyme F420-6 +0.00001Coenzyme F420-7 +0.219Tetrahydrosarcinapterin +0.044Adenosylcobalamin +0.019430Coenzyme F +0.00464Coenzyme B +0.00093THF +0.00005Coenzyme F390a +0.00005Coenzyme F390g +0.191Methanofuran
Carbohydrates	0.01		1.27Glycogen + 0.06Galactan + 2.49Polyacetylgalactosamine + 1.6Polyglucuronate
Osmolytes	1		0.0011Ni^{2+} + 1.11 N-acetyl-beta-lysine + 0.0038Ca^{2+} + 0.4K^+ + 0.0035Fe^{2+} + 0.011Zn^{2+} + 0.001Co^{2+} + 0.163Mg^{2+} + 0.017Na^+

*i*MB745 originally had the SepRS reaction only, leaving this pathway incomplete. The SepCysS reaction was added to the model and connected to the canonical cysteine aminoacylation reaction to complete the alternate pathway (Fig. 3.19). It is important to note that, to our knowledge, the actual sulfur source in the SepCysS reaction remains unknown. However, it has been shown that sodium sulfide provides the highest activity *in vitro* [231]. We decided to use hydrogen sulfide as the sulfur source because it is a sulfide produced within the organism. The parsimonius FBA solution of the model with this modification interestingly showed that it only uses the canonical charging pathway even if both pathways are turned on and will only use the alternate pathway if the canonical pathway is knocked out. The RPKM values from the RNAseq data, however, suggests that SepRS is expressed at least twice as much as CysRS on average across MeOH, Acetate, and TMA (Table 3.20). In order to constrain the pathways to reflect this, flux variability analysis was first used to determine the allowable flux ranges for SepRS and CysRS. The maximum allowable flux through CysRS was then set to equal half that of SepRS. In *i*MB745, allowing any cysteine uptake led to an overproduction of ATP and resulted in unrealistically high growth rates. To reflect media component consumption accurately in addition to the SepRS/CysRS constraints, *i*ST807 constrains the cysteine uptake at a maximum rate of 0.35 $\frac{mmol}{gDWh}$ which is the uptake rate that minimized the differences between simulated and experimental growth rate on methanol, acetate, and carbon monoxide. Dynamic flux balance analysis was also performed to verify that this cysteine is not growth-limiting over a period of about 38 hrs as consistent

with experiment [16]. These constraints successfully forced flux through both cysteine aminoacylation pathways.

Pyrrolysine Biosynthesis Pathway MMA methyltransferase is responsible for activating MMA for a methyl transfer to a cognate corrinoid protetin. In 2002, the crystal structure [232] of this enzyme from *M. barkeri* revealed the presence of pyrrolysine (pyl) within the catalytic site. Later studies by Mahapatra, et al. [233] on *M. acetivorans* showed that this methanogen could not grow on any methylamine substrates (MMA, DMA, and TMA) without the gene for pyl-tRNA, demonstrating that pyl is required for growth on methylamine substrates. The pathways involving pyl synthesis in *i*MB745 were present but flawed. First, the pyl-tRNA charging pathway mistakenly used the alanyl-tRNA instead of the known pyl-tRNA. Second, the hypothesized pyl biosynthesis pathways were outdated and no fluxes ran through them even when successfully simulating the model on methylamine substrates. The model essentially allowed for growth on methylamine substrates without the synthesis of pyl anywhere. These two errors were fixed by replacing the alanyl-tRNA with pyl-tRNA in the pyl aminoacylation pathway and replacing the pyl biosynthesis pathway with the most recent and accepted pathway from Gaston, et al. [234] shown in Fig. 3.21.

In order to force the model to recognize that pyl is required for growth on methylamine substrates, the biomass reaction was modified to reflect amino acid use more accurately. The first modification altered the biomass reaction to draw in aminoacylated tRNAs instead of free amino acids, keeping the

coefficients the same. This change was based on the realization that it is amino acids charged onto their tRNAs that eventually become part of the cell biomass through protein synthesis rather than any free amino acid produced in the cell. The second modification adds a pyl-tRNA term to the biomass when simulating growth on methylamine substrates. For non-methylamine growth, the model sets the coefficient for pyl-tRNA to zero. For methylamine growth, the model turns on the coefficient. This coefficient was estimated from the approximate number of CTA codons in the *M. acetivorans* genome. CTA is canonically a stop codon but regulatory mechanisms exist within *M. acetivorans* to express this as the pyl amino acid. Since the fraction of CTA codons actually coding for pyl is unknown in this methanogen, it was taken to be 50% as an upper limit estimate. This gives a pyl-tRNA biomass coefficient of 0.081 $\frac{mmol\ pyl}{gDW protein}$.

Biosynthesis Pathway During the curation of *i*MB745, the methanofuran biosynthesis pathway for *M. acetivorans* was largely hypothesized based on the known pathway from *M. janaschii* at the time. Since then, concrete experimental evidence documenting gene-reaction associations for the methanofuran biosynthesis pathway in *M. janaschii* has been published by the White lab [201–204]. The homologous genes *MA4436*, *MA0636*, and *MA1475* in *M. acetivorans* were identified and respectively incorporated into the model for the reactions MFRS1, MFRS2, and MFRS3. MFRS6 and MFRS7 were deleted from the model to match the experimentally verified pathway in *M. janaschii*.

Adding Osmolytes to Biomass Reaction In 1995, Sowers and Gunsalus [208] published a study in which they measured the concentrations of unbound cations within *Methanosarcina spp.* in media with varying osmolarity. It was found that Mg^{2+}, Na^{+}, B, Zn^{2+}, Ca^{2+}, Fe^{2+}, Ni^{2+}, and Co^{2+} concentrations remained relatively constant despite the changing extracellular osmolarities. These unbound cations were incorporated into the biomass reaction with the coefficients: K^{+}, 0.4; Mg^{2+}, 0.163; Na^{+}, 0.017; B, 0.012; Zn^{2+}, 0.011; Ca^{2+}, 0.0038; Fe^{2+}, 0.0035; Ni^{2+}, 0.011; and Co^{2+}, 0.001 mmol/gDW. The osmolyte N-epsilon-acetyl-beta-lysine was included in a the biomass expression at a ratio of 1.11 mmol/gDW.

Adding Gluconeogenesis Intermediates/Products to Biomass Reaction A recent paper measured glycogen, gluconeogenesis fluxes and gluconeogenesis intermediate concentrations in *M. acetivorans* growing on methanol in exponential growth and stationary phase [209]. The glycogen content was significantly higher than assumed in the *i*MB745 model. As such, we have added/updated the biomass coefficients for glycogen and these intermediates based on these new quantitative measurements. The high glycogen content ($\sim$ 0.93651mmol/gDCW), consumes significant energy of the cell during growth; therefore, the ATP maintenance cost had to be lowered to match growth experiments. A final value of 44.1 mmolATP/gDCW. Comparing this value to the previous value of 65.0 mmolATP/gDCW indicates that nearly 33% of energy derived by the methanogen is used in storing glycogen. This could confer evolutionary advantage when nutrients are scarce.

Modeling Alternate Biomasses for Different Growth Substrates

Biomass for growth on acetate and TMA were fit taking MeOH to be associated with the published biomass coefficients. Additionally, acetate was fit using TMA as the starting biomass coefficients. Fit biomass coefficients can be seen in Figure 3.26. Flux comparisons for MeOH vs Acetate and MeOH vs TMA can be seen in Figs. 3.32 and 3.33 respectively, where it is demonstrated that significant changes to fluxes in amino acid and cofactor biosynthetic pathways are predicted. Many coefficients can vary significantly, as indicated by the large standard deviations and it is unclear as to whether physiologically requirements differ. However, a handful of biomass components were statistically different ($p<0.01$, t-test, n=96) in the second condition compared with the first, possibly suggesting a different physiological requirement (a greater or smaller fraction of total biomass when growing in one media compared to another). When comparing methylotrophic to acetotrophic growth, our fitting procedure suggests that nickel, cobalt and AMP requirements decreases, while cellular zinc and potassium increase. The cobalamin cofactors in methyltransferases contain cobalt and the downregulation of these enzymes under acetotrophic growth is consistent with a decrease in requirement for cobalt [16]. The decreased requirement for nickel is counterintuitive as methyl-coenzyme M reductase and carbon monoxide dehydrogenase are both nickel containing and upregulated on acetate growth. Decrease in AMP requirement could be due to slower growth relating to the phosphate balance.

The average coenzyme M (CoM) increases by a factor of 10 going from

MeOH to acetate, similar to results from a recent paper that showed CoM, and sulfide content in general, increases roughly by a factor of 2.8–3x for acetate grown cells [224]. Interestingly, glycogen galactan and polyglucuronate are predicted to be produced at higher levels, along with phosphorylated *myo*-inositol phospholipids, which are possibly used in producing extracellular matrices that are common in cell aggregates for acetate grown *M. acetivorans* [195].

The fitting procedure suggests a decreased biomass requirement for adenosylcobalamin and coenzyme B (CoB) when grown in acetate as compared to TMA. The former can be explained by the fact that methyltransferase which are composed of at least one corrinoid cofactor are severely down-regulated when grown on acetate (and generally not of use to acetotrophic methanogenesis). Most cofactors were required in higher levels for growth on TMA than on MeOH, including coenzyme A, both the adenosinyl- and guanosinyl-coenzyme F390 analogs, succinyl-CoA and tetrahydrofolate. The reason for these increases is unclear.

The modeling indicated that several variants of coenzyme F420 was generally lower while grown on MeOH than on acetate or TMA, however the reason for this is unclear. Reports in literature on various *M. barkeri* and *M. mazei* strains are conflicting; one study indicating that F420 concentrations are significantly higher comparing MeOH to acetate grown cells [240], while other studies found a higher level for acetate grown cells [136, 241, 242]. Interestingly, one report found that different variants of the coenzyme F420 predominated in different *Methanosarcina* species. Further experiment and

modeling is required to uncover the role of the various analogs and their regulation.

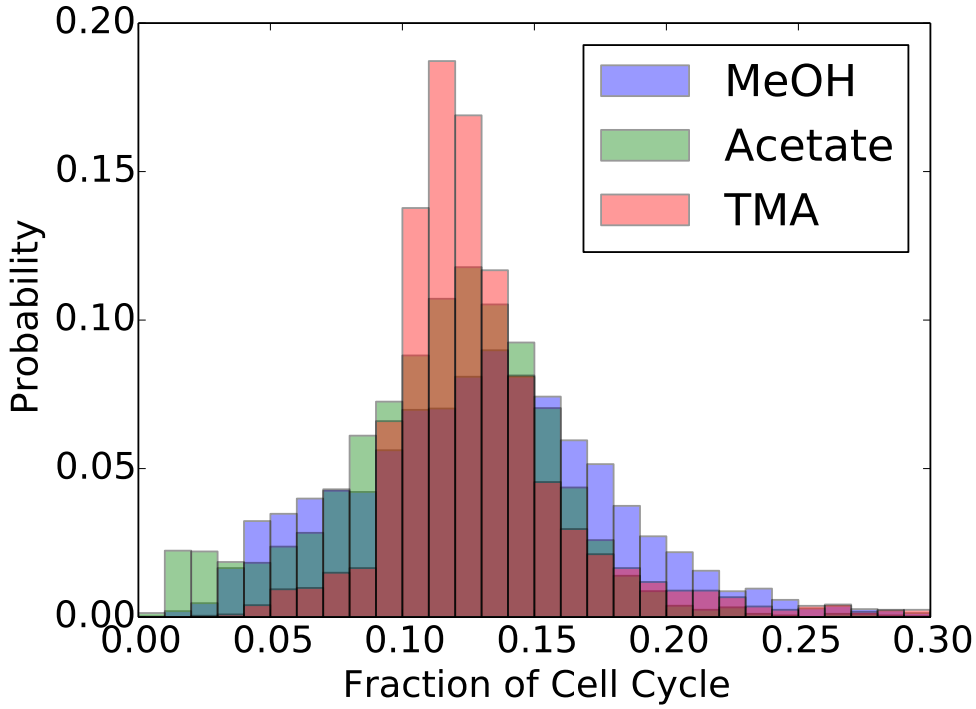


Figure 3.15: **Normalize Half-Life Distributions.** The distributions that result after the half-lives have been normalized by doubling time in the respective condition. The scaled half-life is a measure of the fraction of the cell cycle that an RNA molecule is likely to persist. As can be seen, the scaled distributions overlap and the mean fraction of the cell cycle that an RNA persists (0.127 ± 0.035) are not statistically different ($p > 0.33$, t-test.)

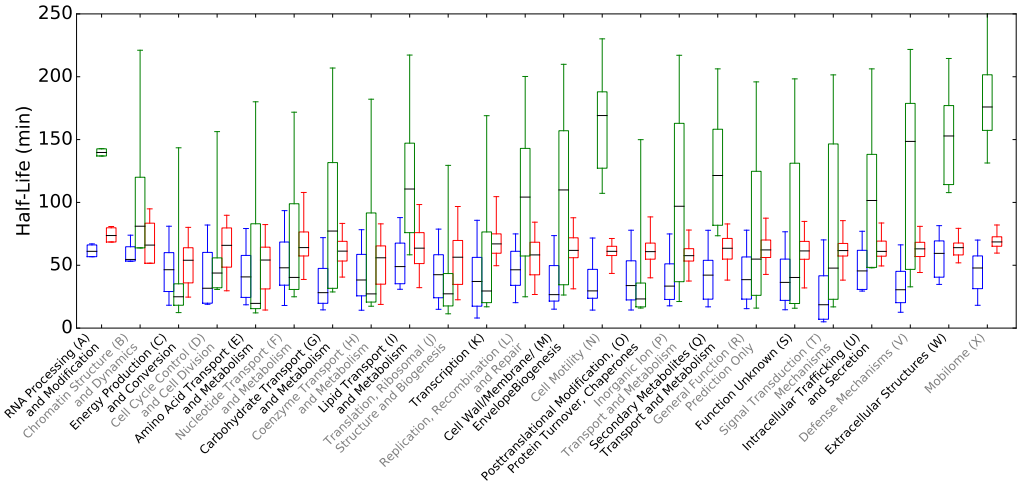
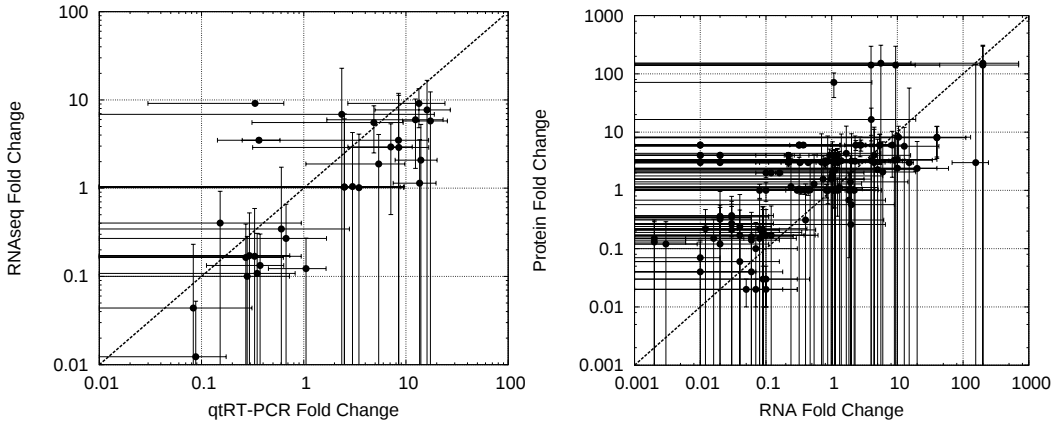


Figure 3.16: **mRNA Half-Life Statistics by Class.** mRNA half-life statistics by class showing median and quartiles for MeOH (blue), acetate (green) and TMA (red) growth. Percentiles were computed using the weighting method of Edgeworth [226]. The overall range (whiskers) of the distributions are generally the same across classes, however the quartiles and median can be significantly different, supporting the conclusions in the main text that mRNAs are selectively stabilized/destabilized depending on function.



(a) (b)

Figure 3.17: **Comparison of RNAseq Data to Previous Experiments.** The line indicates the exact diagonal. (A) A comparison of fold change between conditions computed from our RNAseq data of this study to qtRT-PCR or Microarray data from previous studies shows a linear relationship with a slope of 0.96 and an overall correlation of 0.82. (B) A comparison of fold change from our RNAseq data to fold change in reported protein abundances demonstrates a correlation of 0.63.

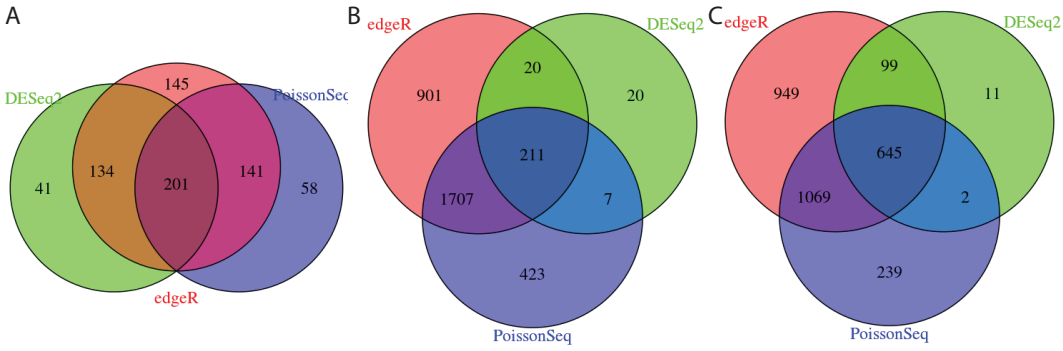


Figure 3.18: **Overlap of DEG Calling Methods.** Count of differentially expressed genes where $p \leq 0.01$ predicted by each method when comparing: (A) Methanol vs. Acetate, (B) Methanol vs. TMA, and (C) TMA vs. Acetate. In general DESeq2 is the most conservative method. The overlap drastically reduces the number of DEG and provides a more certain set of predictions.

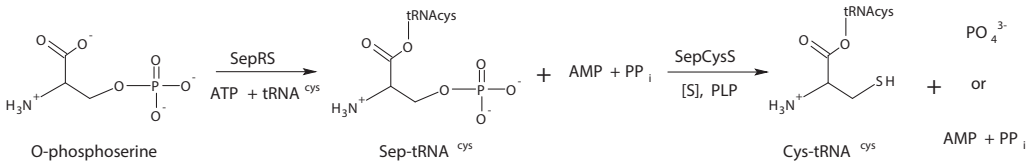


Figure 3.19: **Cysteine Biosynthesis.** Alternate cysteine aminoacylation pathway involving SepRS and SepCysS enzymes.

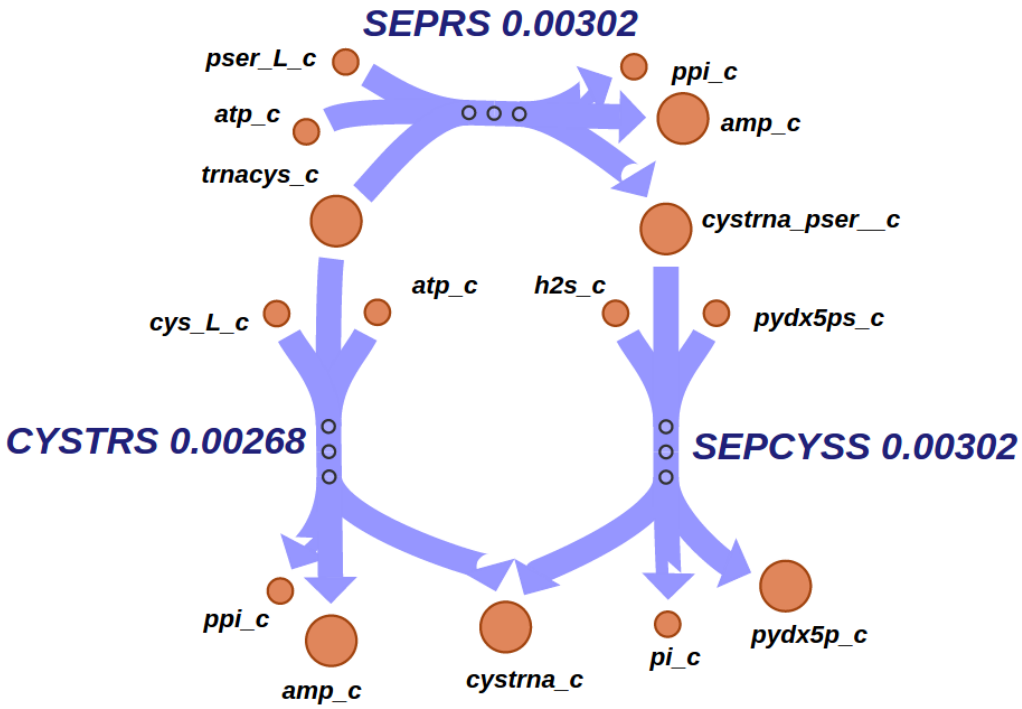


Figure 3.20: **Cysteine Pathway Fluxes.** Flux distribution of CysRS and SepRS under MeOH growth.

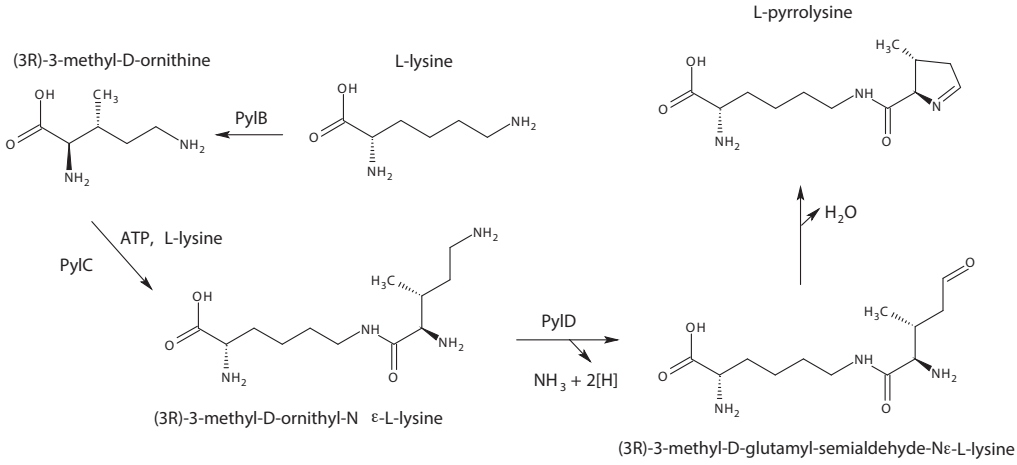
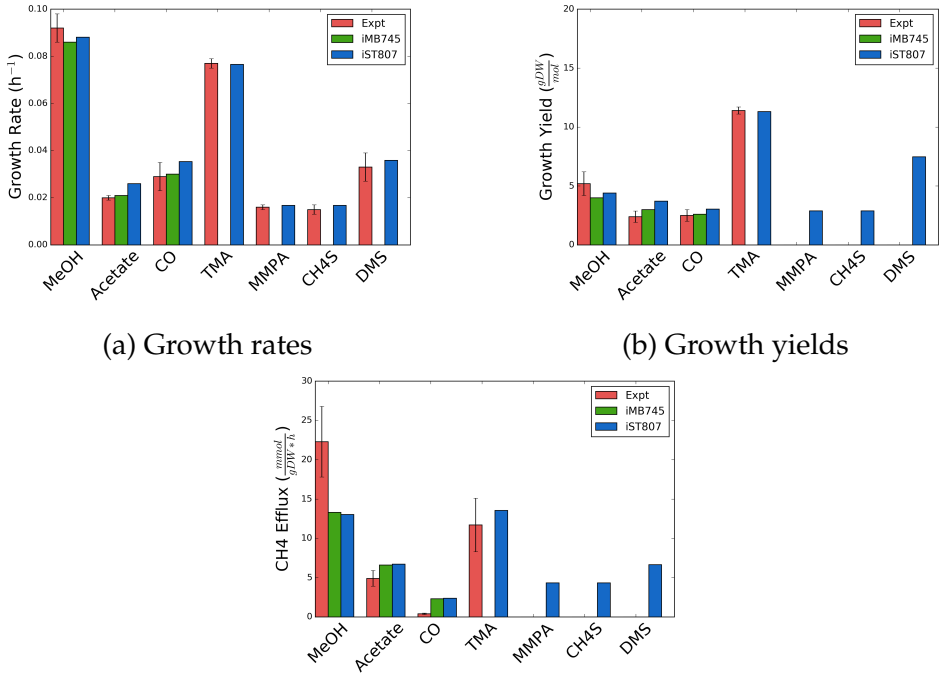


Figure 3.21: **Pyrrolysine Biosynthesis Pathway.**



(a) Growth rates

(b) Growth yields

(c) CH_4 production

Figure 3.22: **Model Predictions Compared with Experimental Data.** A) Growth rates (hr^{-1}) [17, 19, 23, 25, 104, 134, 143, 194, 195, 217, 235]. B) Growth yields (gDW/mol substrate) [195, 236, 237]. Note that experimental TMA growth yield was computed from TMA growth rate and a fitted TMA uptake rate as there were no experimental uptake values available. C) CH_4 production rate (mmol/hr/gDW) [104, 237–239]

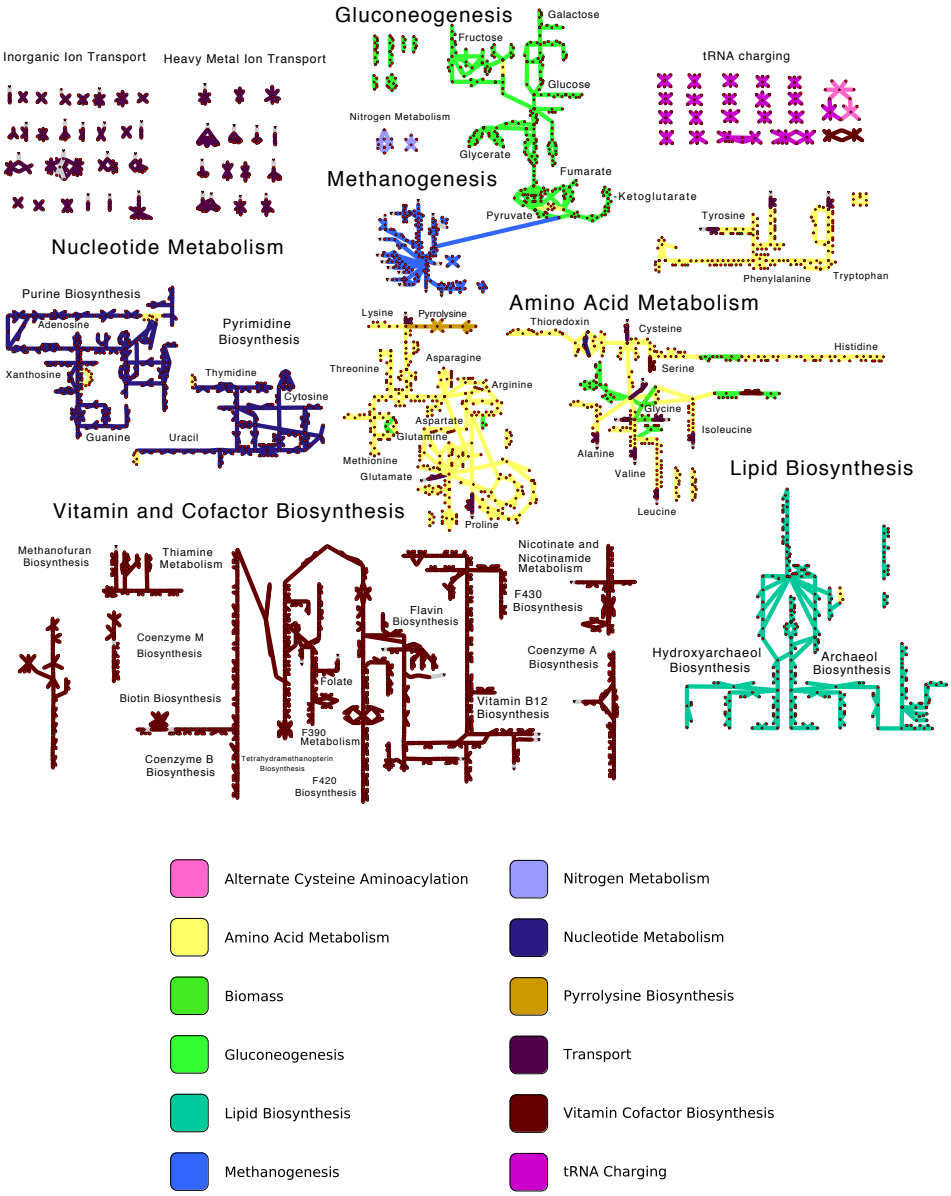
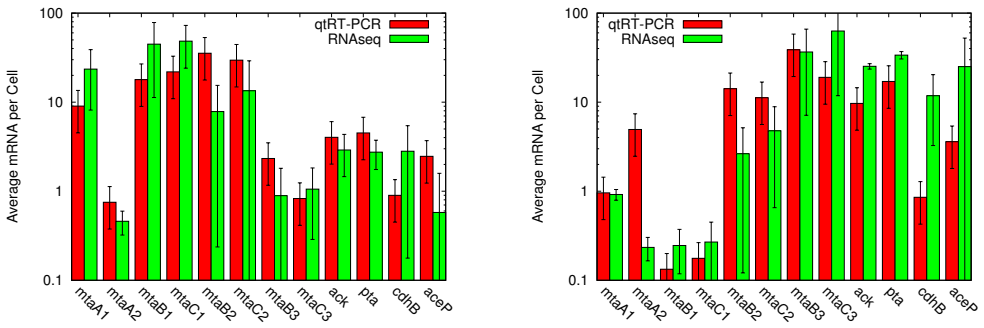


Figure 3.23: ***M. acetivorans* Metabolic Map.** Metabolic map of the updated *M. acetivorans* model published with this paper. Red dots denote reactions and metabolites while edges describe the connection of metabolites through reactions. In general, reactions are drawn such that metabolic flux flows down and to the right; however, there are many exceptions notably in gluconeogenesis and cofactor/lipid metabolism. Reactions are colored by metabolic classification. This map is available in format in formats compatible with Cytoscape [210] and Escher [211] in the Supporting File "ModelAndMaps.zip".



(a) (b)

Figure 3.24: **mRNA Copies Per Cell.** A comparison of mRNA copies per cell estimated via our RNAseq data, and previous studies that utilized RT-qPCR to quantify transcript abundance in the related organism *Methanosarcina mazei* [212] grown in (A) methanol and (B) acetate. Error bars are standard deviation of the mean for 3 replicates. Values from Cao et al. are for cells grown at 30°C compared to our cells which were grown at 37°C. All values agree within uncertanties except for *cdh*, *mtaA2*, and *mtaB2* indicating the organisms have similar expression profiles and our estimates for mRNA counts are good.

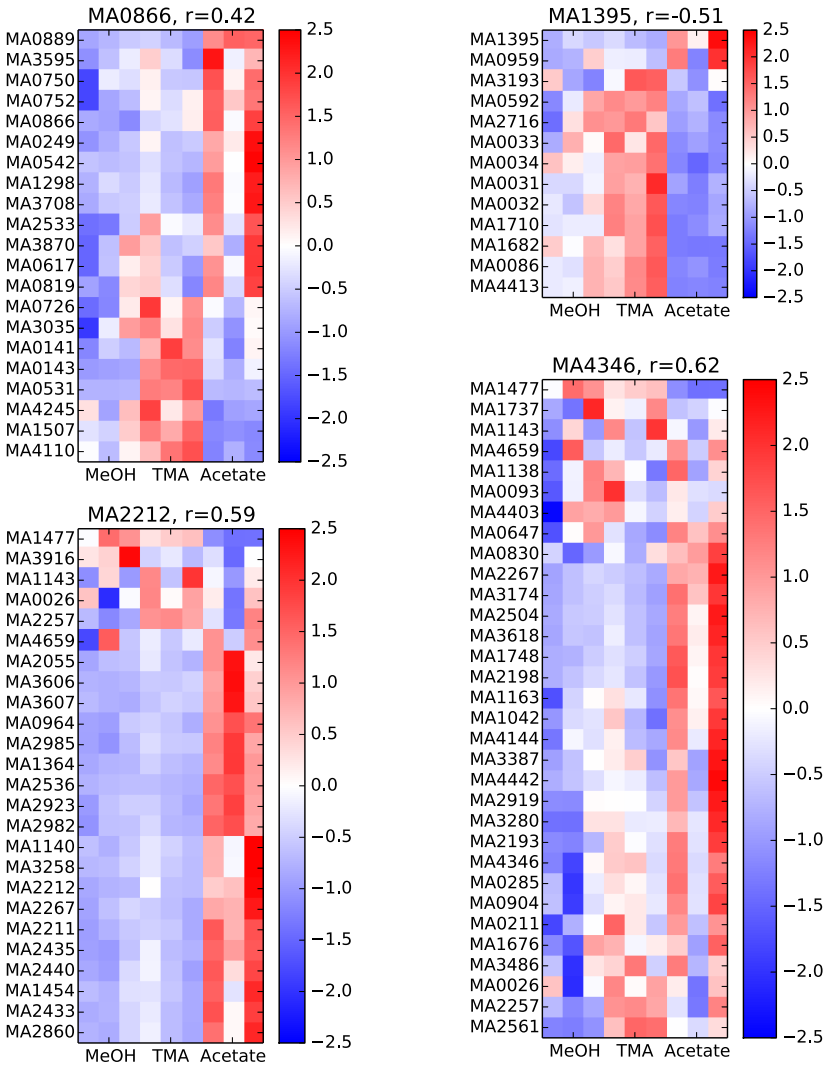


Figure 3.25: **Genes Correlated with Transcription Factors.** Heatmaps of relative expression for 4 putative regulators that are differentially expressed between at least one pair of conditions. Each regulator is highly conserved among the *Methanosarinales*. *MA1395* is highly conserved among most methanogens and encodes for a nickel response regulator. The regulator is indicated in the title of each heatmap along with the correlation of the regulator's expression to the other genes that have the same conservation pattern.

Figure 3.26: **Sampled Biomass Coefficients.** Biomass coefficients after fitting metabolic flux distributions to (green circles) compared with biomass coefficients published with the original model (orange squares). In each case, the substrate listed first was taken to be the reference, and the differentially expressed genes going to the second substrate were used to fit the flux distributions and subsequently the biomass composition (see SI Section 3.6.3). Results shown for: (a) Methanol vs. Acetate, (b) Methanol vs. TMA, and (c) TMA vs. Acetate.

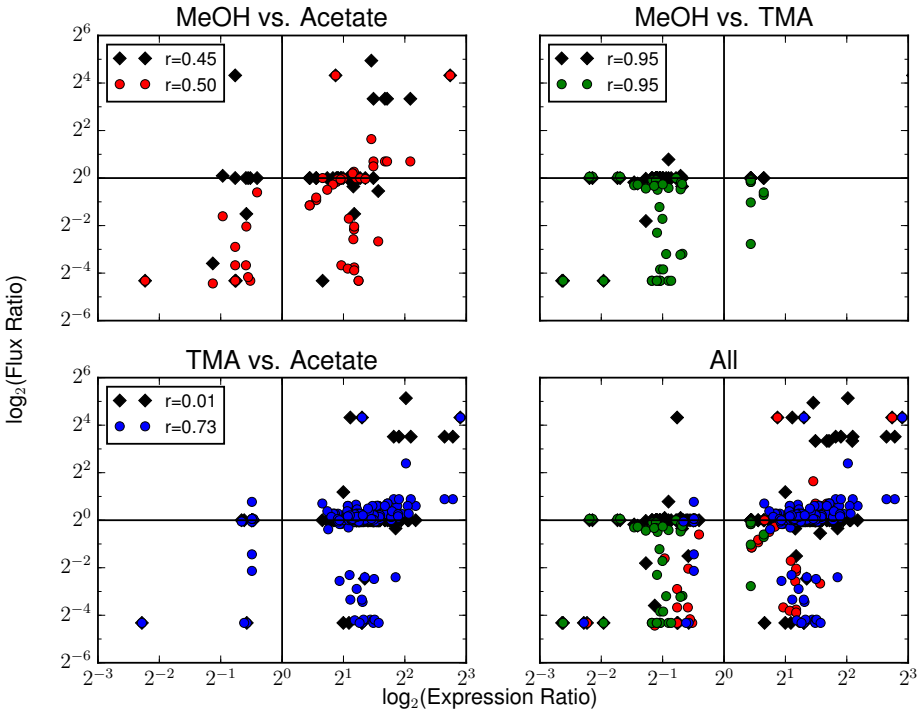


Figure 3.27: **Improved Flux Predictions.** A comparison of the ratio of fluxes of the first to second substrate computed over the whole metabolic model, to the ratio of DEG for that reaction. Predictions pre-fitting are shown as black diamonds while values after fitting are shown as circles. Overall, correlation to the experiment increases from r=0.33 to r=0.51.

Figure 3.28: **Conservation of Genes; MeOH vs TMA.** Conservation of the genes that are differentially expressed between MeOH and TMA across the tree of methanogens. Each vertical bar indicates that a homolog for the differentially expressed gene exists in the indicated species (computed as the bidirectional best hits functionality in the ITEP software [191] with an E-value cut-off of 10^{-5} for a database of $\sim$125000 proteins). Most differentially expressed genes are highly conserved among the *Methanosarcinales;* however a core set of genes are conserved across all methanogens.

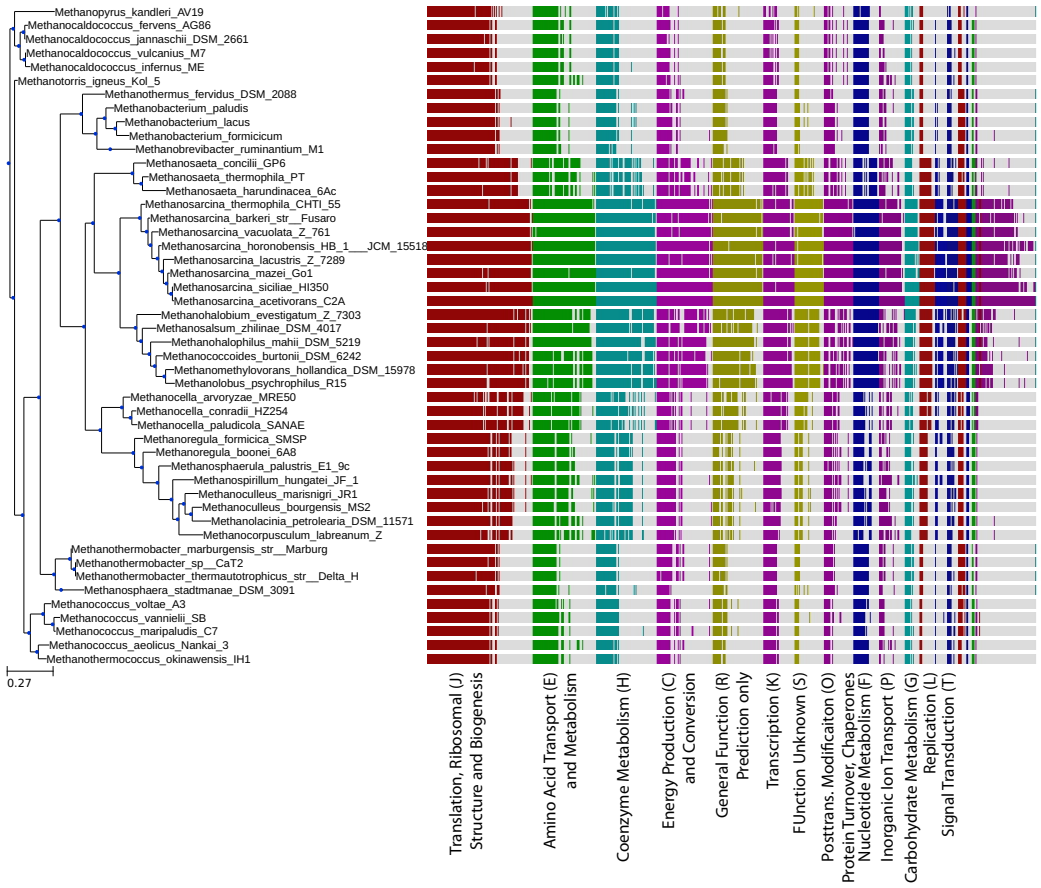


Figure 3.29: **Conservation of Genes; TMA vs Acetate.** Conservation of the genes that are differentially expressed between TMA and Acetate across the tree of methanogens. Each vertical bar indicates that a homolog for the differentially expressed gene exists in the indicated species (computed as the bidirectional best hits functionality in the ITEP software [191] with an E-value cut-off of 10^{-5} for a database of $\sim$125000 proteins). Most differentially expressed genes are highly conserved among the *Methanosarcinales;* however a core set of genes are conserved across all methanogens.

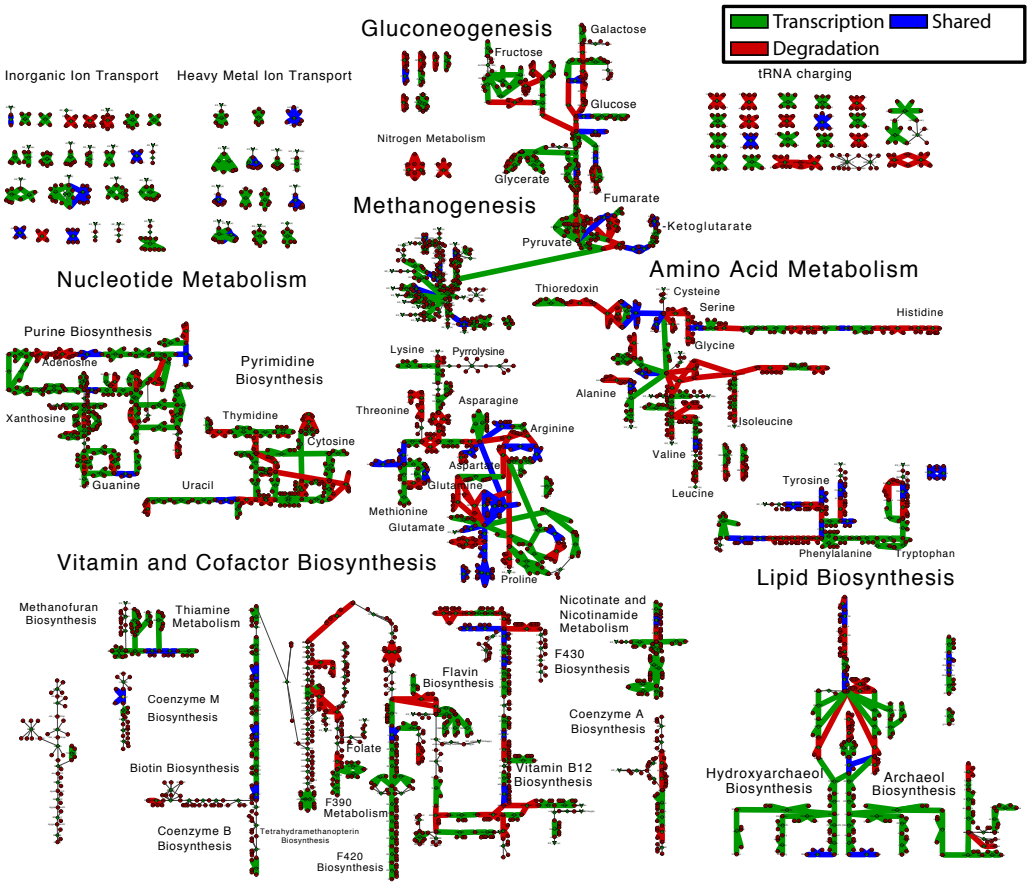


Figure 3.30: **Control Coefficient Map; MeOH vs TMA.** A mapping of the control coefficients for changing mRNA expression levels between MeOH and TMA. Red indicate reactions where mRNA levels are regulated by shifts in the degradation rate, while green indicates mRNA level shifts due to changes in transcription rate. Blue indicates reactions where mRNA levels are affected by both transcription and degradation rate.

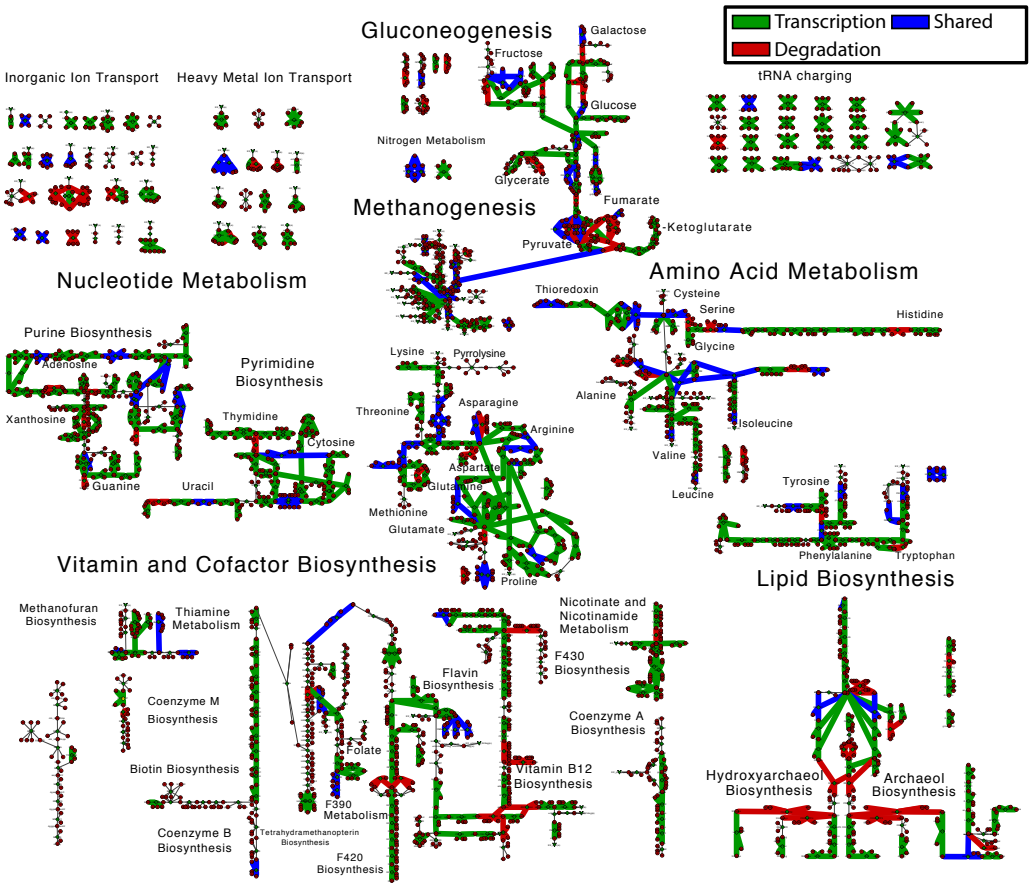


Figure 3.31: **Control Coefficient Map; TMA vs Acetate.** A mapping of the control coefficients for changing mRNA expression levels between TMA and acetate. Red indicate reactions where mRNA levels are regulated by shifts in the degradation rate, while green indicates mRNA level shifts due to changes in transcription rate. Blue indicates reactions where mRNA levels are affected by both transcription and degradation rate.

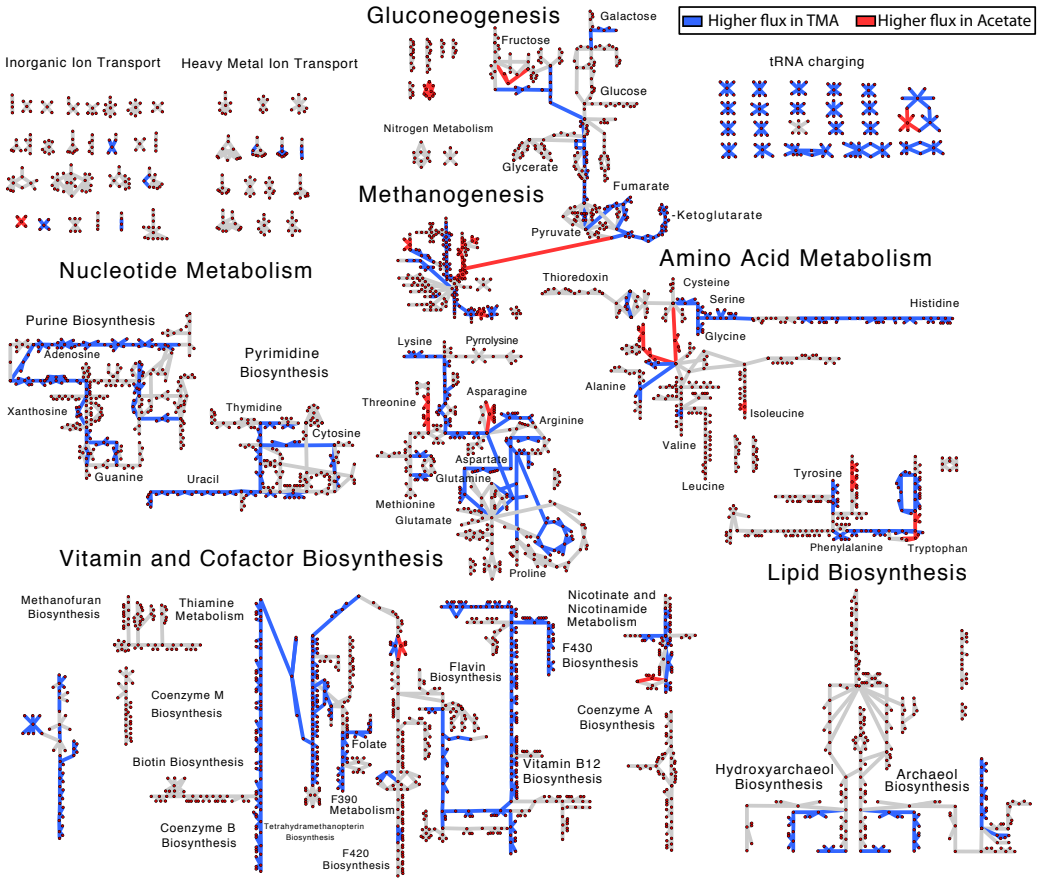


Figure 3.32: **Metabolic Fluxes; MeOH vs Acetate.** Changes in metabolic pathway usage that is consistent with the differentially expressed genes comparing MeOH to acetate growth. Pathways with significant changes in fluxes are shown in red (up in acetate) and cyan (down in acetate) while reactions showing no change in flux (change in flux <2x) or having no associated genes in gray. Significant metabolic changes are observed across nearly all of metabolism, however no changes are are predicted for coenzyme A and M biosynthesis, thiamine metabolism, or leucine/isoleucine/methionine synthesis. Additionally, only changes in phosphoethanolamine and phosphoglycerol based lipids.

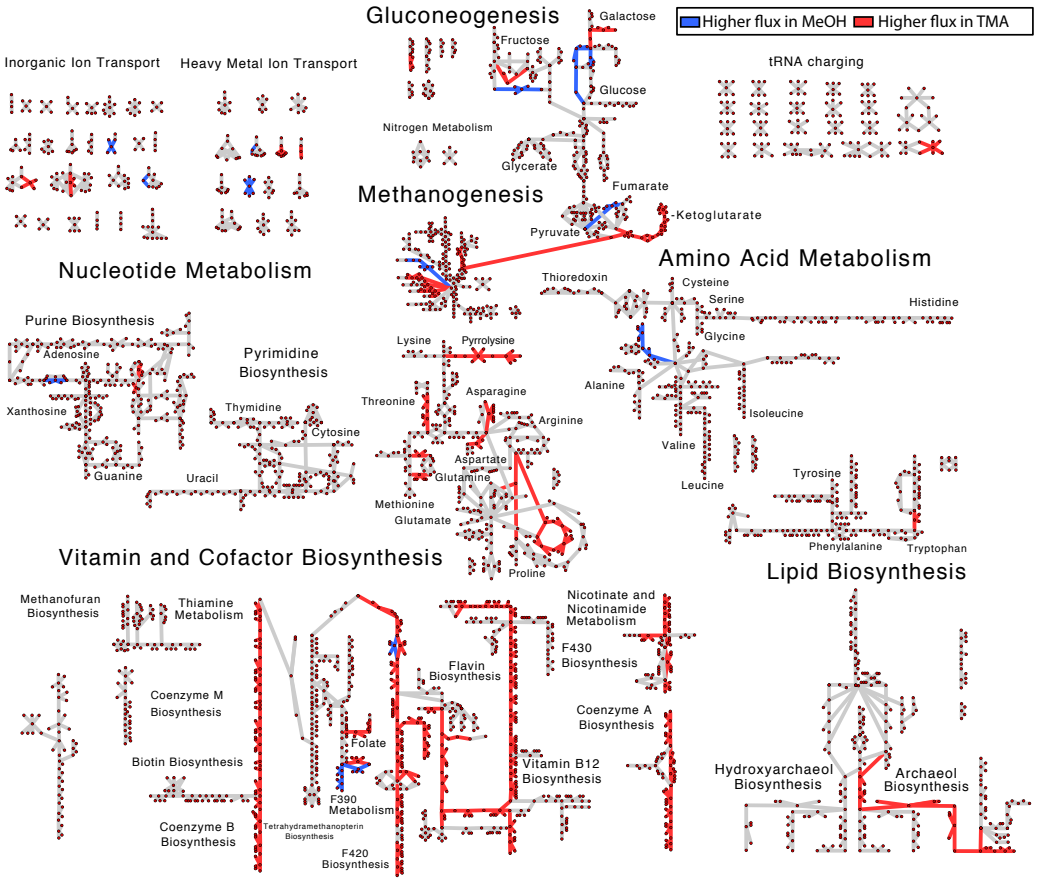


Figure 3.33: **Metabolic Fluxes; MeOH vs Acetate.** Changes in metabolic pathway usage that is consistent with the differentially expressed genes comparing MeOH to TMA growth. Pathways with significant changes in fluxes are shown in red (up in TMA) and cyan (down in TMA) while reactions showing no change in flux (change in flux $<$2x) or having no associated genes in gray. Cofactor and vitamin metabolism show the most significant changes, along with central amino acid metabolism (including production of α-ketoglutarate and malate) and the pathway producing glucosaminyl archaetidyl-myo-inositol lipids.

Chapter 4

Genome-Scale Metabolic Modeling of Archaea Lends Insight into Diversity of Metabolic Function

Decades of biochemical, bioinformatic and sequencing data are currently being systematically compiled into genome-scale metabolic reconstructions (GEMs). Such reconstructions are knowledge-bases useful for engineering, modeling and comparative analysis. Here we review the fifteen GEMs of archaeal species that have been constructed to date. They represent primarily members of the Euryarchaeota with almost three quarters representative of methanogens. Unlike other reviews on GEMs, we specifically focus on archaea briefly reviewing the construction process and the genealogy of archaeal models. The major insights gained during the construction of these models are then reviewed with specific attention to novel metabolic pathway predictions and growth characteristics. Metabolic pathway usage is discussed in the context of the composition of each organism's biomass and their specific energy and growth requirements. We then show how the metabolic models can be used to study the evolution of metabolism in archaea. Conservation of particular metabolic pathways can be studied by

The contents of this chapter are based in part on work previously published as Sheng-Shee Thor, Joseph R. Peterson and Zaida Luthey-Schulten. "Genome-Scale Metabolic Modeling of Archaea Lends Insight into Diversity of Metabolic Function," *Archaea*, vol. 2017, Article ID 9763848, 18 pages (2017) [37]. Specifically, S.T. worked closely with me to analyze the methanogens and created figures 4.2-4.4.

comparing reactions using the genes associated with their enzymes. This demonstrates the utility of GEMs in evolutionary studies, far beyond their original purpose of metabolic modeling; however, much needs to be done before archaeal models are as extensively complete as those for bacteria.

4.1 Introduction

Since their discovery and classification in the late 1970s and early 1980s [1, 243–246] archaea have garnered considerable interest due in part to prevailing thoughts at the time that they lived primarily in extreme conditions, a property that results in unique cell physiology and metabolic characteristics [247]. Although the original classification of organisms was based on only thirteen sequences with only four representatives of archaea [1], the proposal of the three domains of life has been tested time and time again [82, 87, 247–249] and holds up remarkably well. Archaea have now been found to reside in essentially every terrestrial environment, but the observation of their unique physiological and metabolic properties still holds [247]. They can produce and consume methane, reduce or oxidize sulfur and iron containing compounds, and perform either nitrification or denitrification [250, 251].

Despite the significant progress in sequencing archaeal genomes, a systematic understanding of the metabolism of Archaea is still lacking. This is especially true for peripheral metabolic pathways and mechanisms of adaptation to extreme environments. It has often been noted that the en-

vironmental niches dominated by Archaea constitute extremely stressful or even fatal homes for their bacterial cousins; thus, they have evolved unique coping mechanisms and optimized their metabolisms to salvage the energy that would otherwise be left unused in the environment. It has been proposed that adaptation to energy stress is the primary factor driving the evolution of Archaea [252]. The consequence is that they have evolved specialized tolerance and metabolic capabilities unique to their environments which makes them relatively inflexible to adaptation like their bacterial counterparts. It has been proposed that this inflexibility results in tighter phylogenetic groups that directly represent less metabolic diversity [252]. Indeed, the evidence seems to support this hypothesis as only 89 genera of archaea have been identified in contrast to the over 1,400 bacterial genera. This fact should be exploitable by systems biology researchers as it means that information gained by one member of a phyla can largely be extended to other members of the phyla.

For this reason, systematic databases of the metabolic properties of the Archaea are highly desirable; the field of systems biology is uniquely positioned to provide useful insight into the diversity and evolution of metabolic capabilities. To date, fifteen genome-scale metabolic models (GEMs; one of the main products of systems biology research) have been constructed for ten archaeal species. However, these models represent primarily members of the Euryarchaeota with almost three quarters representatives of methanogens. An examination of the phylogenetic tree demonstrates a lack of well curated metabolic reconstructions in many of the Archaea phyla (see Fig. 4.1).

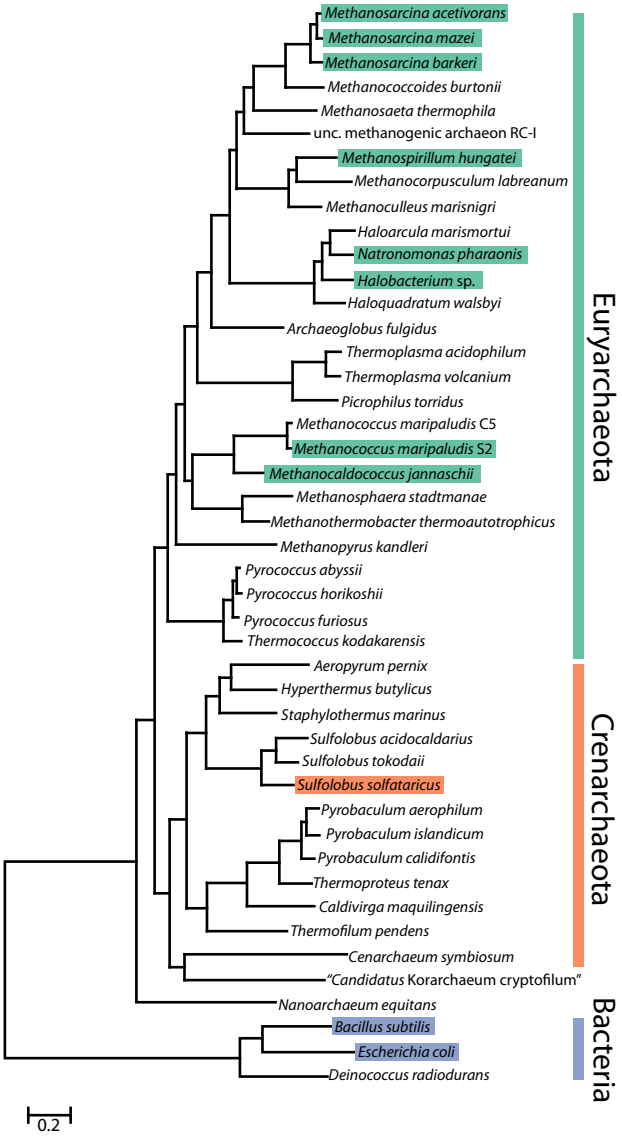


Figure 4.1: **Diversity of Achaeal Models.** A visualization of the diversity of archaeal genome-scale metabolic models as related by phylogeny. The figure is adapted from Elkins, et al. [253] where the maximum-likelihood tree was constructed using 33 conserved ribosomal proteins and three three largest RNA polymerase subunits. Highligthed species indicate that genome-scale metabolic models have been constructed for that organism. Although not shown in this adapted figure, a *Methanobrevibacter smithii* model within the Euryarchaeota has also been constructed. While numerous models have been constructed for Euryarchaeota, the Crenarchaeota are severely under-represented.

Despite the limited representation, much can already be learned from the GEM "knowledge-bases". Here we review the GEMs constructed to date and the knowledge gleamed from them. We begin by briefly reviewing the construction and predictions of these metabolic models. We also give a historical perspective on the construction of archaeal GEMs comparing their various states of completeness. We then review the models and specific insights gained from model constructions, including novel metabolic enzymes/pathways. A section is devoted to methanogens—as the most heavily studied Archaea—with an analysis of the phylogeny and performance of the metabolic models. Additionally, we describe the insights learned during model construction for non-methanogenic archaea especially with regard to the novel metabolic pathways and growth characteristics. Finally, we demonstrate the utility of these metabolic models to the study of evolution of diversity in Archaea. We do this by computing the conservation of reactions (based on genetic association of the enzymes) across the Archaea and visualizing it a comprehensive map of the metabolism of the methanogen *Methanosarcina acetivorans*. The results demonstrate that in general amino acid metabolism is highly conserved and other interesting observations about proline, central, and nitrogen metabolism.

4.2 Genome-Scale Metabolic Models (GEMs)

Metabolic networks are invaluable visualization tools for qualitatively understanding an organism's metabolic behavior under given conditions and

have a long history of use in biology. Systematic construction of metabolic models—which couple metabolic networks with genetic associations, reactions that exchange metabolites with the environment and the organism's biomass composition—only began to take shape in the mid-1990s when Fleischmann, et al. [254] fully sequenced the entire genome of the bacterium *Haemophilus influenzae* Rd. They compared *E. coli* proteins known at the time against the *H. influenzae* Rd genome and showed that 68% of the known *E. coli* proteins had homologs in the *H. influenzae* Rd genome, enabling them to hypothesize the metabolic pathways that exist in *H. influenzae* Rd. Since then, the pioneering work of Bernhard Palsson and co-workers [61] have established genome-scale metabolic models (GEMs) as the standard computational tool with which to quantitatively study the metabolic behaviors of organisms. In 2010, a well-established workflow was published in an article detailing the best practices for the model construction process [61].

A GEM can best be described as a knowledge-base containing all the biochemical information describing an organism's metabolic network. They are typically presented as Systems Biology Markup Language (SBML) [255] files that can be queried to obtain information about individual reactions, metabolites, and genes coding for the enzymes that catalyze the reactions. Software such as MATLAB COBRA Toolbox [256, 257] or COBRApy [230] that implement Constraint-Based Reconstruction and Analysis (COBRA) methods can then use the information within a GEM to compute predicted metabolic behaviors of the organism [69]. Alternatively, one can create independent analysis tools that simply use GEMs to identify product synthesis

pathways [258–261], optimize bioprocessing efficiency [262, 263], predict metabolic engineering targets [263, 264] and elucidate more complex phenomena such as microbial communities [265–269].

4.2.1 Model Construction and Predictions

Here, we will briefly summarize the GEM construction process and highlight the most important characteristics of GEMs one typically encounters. The *de facto* standard GEM construction protocol is that published by Thiele and Palsson [61] and should be referred to for standards within the field.

The construction process proposed in [61] is divided into four broad stages: (1) automated construction of a draft model, (2) manual refinement of the draft model, (3) conversion of the model into a mathematical model, and (4) quantitative evaluation and refinement of the model. The first stage involves identifying all the potential reactions and pathways that the organism harbors based on its annotated genome. This process can be automated as it is essentially a bioinformatics problem requiring the comparison of the genome with databases that document known genes and their associated metabolic enzymes and pathways (e.g. KEGG [270], Uniprot [271], BioCyc [272]). Many tools have been designed to facilitate this process such as the RAVEN toolbox [273] and the ModelSEED [274].

The manual curation stage of draft model refinement is the most time-consuming—arguably the most critical—portion of the process. All the reactions and pathways identified in the first stage are evaluated to ensure a variety of consistencies—experimental data should support their existence

in the organism, mass and charges need to be balanced and consistent with reaction stoichiometries, and reaction directionalities need to be consistent with thermodynamic data. Missing pathways are added at this stage along with transport reactions responsible for the organism's intake and expulsion of metabolites. The most crucial features that make GEMs unique and enable subsequent quantitative predictions are also established in this stage; specifically, the biomass objective reaction, the growth-associated ATP maintenance reaction (GAM) and the corresponding non-growth associated ATP maintenance (NGAM) reaction, and the boolean gene-protein-reaction associations (GPRs).

Quantitative prediction using GEMs is typically framed as a linear programming problem in which one feature of the model is optimized under a given set of constraints. This feature is typically the model's biomass production rate which is described by a single pseudo-reaction that produces a "biomass" pseudo-metabolite by drawing in all the metabolites that the organism requires to physically grow (e.g. individual amino acids, carbohydrates, lipids, nucleic acids, vitamins, cofactors, ions and trace metals). Ideally, this reaction is constructed using the experimentally characterized cell biomass composition of the organism. However, this data is often difficult to obtain, leaving curators to either estimate biomass compositions from the organism's genome or adopt the compositions available from other organisms.

The GAM and NGAM reactions consume ATP. The GAM reaction reflects the ATP consumption required for the organism to grow whereas the

NGAM reaction reflects the organism's basal ATP consumption required to survive but not necessarily grow (e.g. maintain membrane potential and redox balance). Since both reactions reflect the organism's energy requirements in the model, the choice of stoichiometric coefficients for these two reactions greatly influences the model's growth predictions. Ideally, the stoichiometric coefficents of these two vital reactions should be determined from a chemostat experiment in which the growth rate is tracked alongside ATP consumption (or some fiducial metabolite tracing ATP consumption). In practice, one will find that researchers often use a variety of estimation schemes based on the experimental data at hand.

The GPRs are boolean expressions containing the genes that code for metabolic enzymes facilitating the reactions. By piecing the genes together in series of AND and OR operations, a GPR encodes which genes are necessary for an enzyme to be synthesized by the cell and therefore which genes are required for a metabolic reaction to exist. Predictions of gene knockout effects are commonly computed with GEMs. Not all reactions in the model will have GPRs due to either the lack of experimental gene characterizations, the use of non-physical "gapfill" reactions [275], or the presence of novel uncharacterised pathways hypothesized by the curator.

Once this manual curation is complete, one can proceed to the third stage of converting the GEM into a quantitatively predictive model. This is done by defining the "objective" reaction to be optimized in the model and constraining the flux ranges on all model reactions. These flux ranges must reflect a specific growth condition to which the organism is subject. During

model construction, most internal reaction fluxes will likely be unbounded, due to the relatively limited biochemical and proteomics data available for most reactions and organisms; it is the transport reaction fluxes that must be constrained to reflect the nutrient availability of the organism's environment. These constraints [276] will have to be applied through COBRA-capable software. Once these model constraints have been set, flux balance analysis can be run to predict the organism's growth rate and the distribution of fluxes through the metabolic network. The fourth stage is validating these predictions with experimental growth data and discrepancies rectified with iterative manual refinement of the model. Numerous tools [273,274,277–282] have been developed over the years to automate many stages of this arduous construction process, allowing researchers to focus their effort on the last stage of model refinement. Dias et al. [277] provides a comparative review of these various computational tools.

4.3 Genealogy of Archaeal GEMs

The genealogy of all the published archaeal GEMs to date is shown in Figure 4.2 (see Table 4.1 for statistics about the various models). The current archaeal GEMs can conveniently be divided between methanogenic and non-methanogenic archaeal species with the former being the most developed due to the ecological roles that methanogens play in the global carbon cycle and their use in wastewater treatment [10]. Although the very first archaeal GEM was developed for *Methanococcus jannaschii* by Tsoka, et al. [283] in 2003,

the majority of the later methanogen GEMs were derived from a model for *Methanosarcina barkeri* (*i*AF692) which was first constructed by Feist, et al. [30] in 2006. This inheritance stems from the fact *i*AF692 was the first manually curated methanogen GEM thoroughly verified against experimental growth data. *M. barkeri* is also one of the most metabolically diverse methanogens in the Euryarchaeota kingdom, capable of consuming acetate, methylamines, methanol, CO, and CO_2/H_2. Two models, *i*VS941 [31] and *i*MB745 [284], were independently constructed for *M. acetivorans* and published in 2011 by the Maranas and Palsson research groups, respectively. *Methanosarcina acetivorans* is equally diverse and similar in metabolism and thus inherited much of the *M. barkeri* GEM characteristics. *i*MB745 was then used as the base model from which to draft the more recent methanogen GEMs, *Methanobrevibacter smithii* (*i*Msi385) [268] and *Methanospirillum hungatei* (*i*Mhu428) [285], both of which were qualified as preliminary reconstruction for use in larger microbial community studies. The most recent *M. acetivorans* models include *i*MAC868 [286] and our own *i*ST807 [36], which are both independent updates to *i*MB745. The GEM for *M. maripaludis* (*i*MM518) was constructed in 2014 by Goyal, et al. [287,288] independent of the other methanogen GEMs. This is not surprising given that by this time GEMs construction had already been well-established.

Non-methanogenic archaeal GEM construction has been largely dominated by the work from Dieter Oesterhelt's research group. In 2008, they released the first manually curated GEM for *Halobacterium salinarum* R-1 (*i*OG478) [290]. This was followed in 2010 by a new GEM for a haloalka-

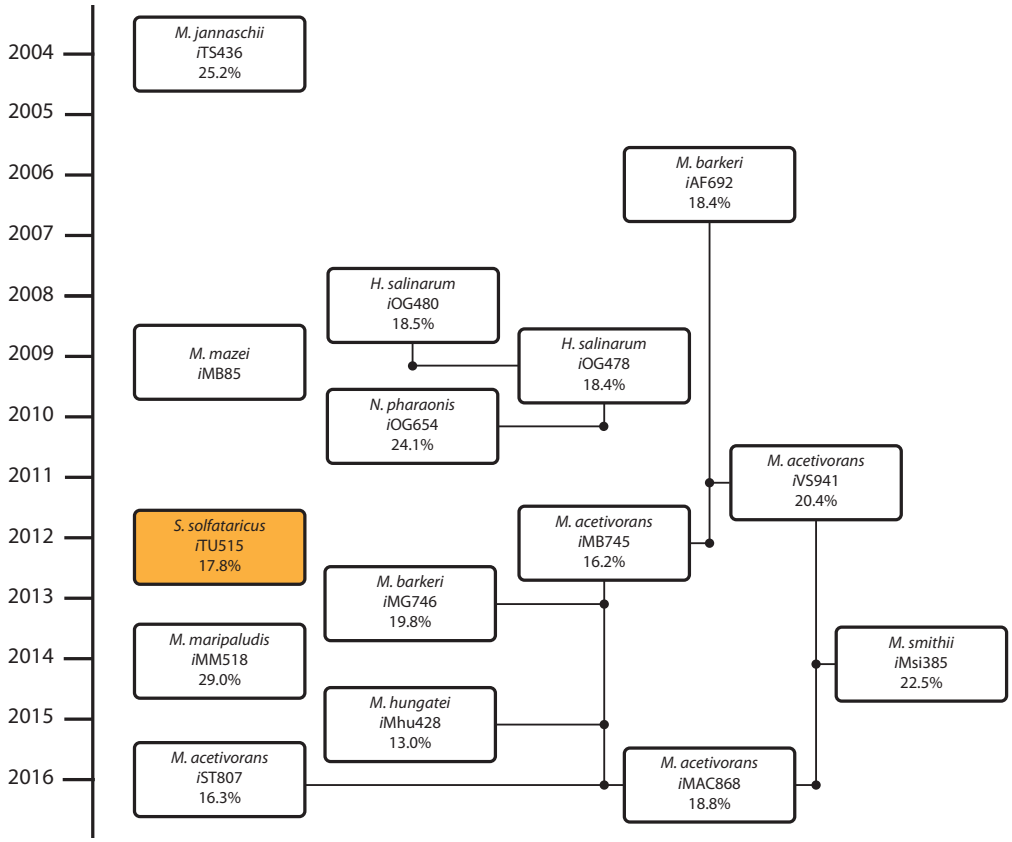


Figure 4.2: **Genealogy of Archaeal Models.** A diagram showing the evolution/genealogy of archaeal models since the reconstruction of *M. jannaschii* in 2004. Each box represents a single metabolic model and includes the species name, the name of the model and the percentage of protein coding genes (where available) that are incorporated in the model. The sole representative of the Crenarchaeota kingdom [289] is highlighted in orange.

Table 4.1: **Model Statistics.** The elementary units of genome-scale metabolic models are genes, metabolites and reactions. These models consist of: 1) a metabolic network that describes the connections between metabolites through reactions (often organized by metabolic subsystem), 2) gene-protein-reaction rules that map the gene dependencies of reactions, and 3) connection to their environment through transport reactions.

Organism	Model	Genes	Metabolites	Reactions	Transport	Subsystems	Citation
H. salinarum	*i*OG490	490	557	711	111	NA	[290]
	*i*OG478	478	545	664	97	NA	[291]
M. acetivorans	*i*VS941	941	708	705	71	60	[31]
	*i*MB745	745	715	818	69	30	[284]
	*i*MAC868	868	707	839	91	31	[286]
	*i*ST807	807	733	759	70	30	[36]
M. barkeri	*i*AF692	692	558	619	88	8	[30]
	*i*MG746	746	718	815	74	31	[292]
M. hungatei	*i*Mhu428	428	639	721	41	29	[285]
M. jannaschii	*i*TS436	436	510	609	1	113	[283]
M. maripaludis	*i*MM518	518	605	570	49	117	[287]
M. mazei	*i*SS85	NA	74	85	5	NA	[293]
M. smithii	*i*Msi385	385	582	525	35	NA	[268]
N. pharaonis	*i*OG654	654	597	683	88	NA	[294]
S. solfataricus	*i*TU515	515	705	718	58	65	[295]

liphile, *Natronomonas pharaonis* (*i*OG654) [294], which inherited significantly from the *H. salinarum* model. The only other non-methanogenic archaeal GEM to our knowledge was independently constructed in 2012 for the *Sulfolobus solfataricus* by Ulas, et al [295].

Although archaea had been established since the 1980s as the third domain of life by the pioneering work of Carl Woese and collaborators [82, 244–246, 249], the lack of experimental data on metabolic characteristics of various archaeal species explains why so few GEMs have been constructed to date. Nevertheless, this early stage of archaeal GEMs development provides a ripe opportunity for the community to grasp the core governing properties of archaeal metabolic networks and perhaps adopt standardized model building practices in order to facilitate more efficient communication of metabolic information among researchers going forward.

4.4 Methanogen GEMs

As the most defining metabolic pathway within methanogens, methanogenesis has been well characterized by numerous biochemical studies over the years. Therefore, the most significant and notable differences between the methanogen models will be found in the methanogenesis pathway and supportive pathways producing novel cofactors for different substrates. The basic framework is shown in Figure 4.3 where CO_2 is reduced to methane in a series of steps. Although this basic framework is well-conserved among methanogens, the key difference lies in the exergonic-endergonic reaction

couplings in the pathway. The first step of CO_2 reduction is an endergonic reaction that oxidizes ferrodoxin and produces formylmethanofuran. In simple hydrogenotrophic methanogens that lack cytochromes, this energy is typically recovered by the methyl-H_4MPT:CoM methyltransferase (Mtr) reaction and the heterodisulfide reductase (Hdr) reaction. Mtr expels Na^+ ions in the process of transferring the methyl group onto coenzyme M (CoM) and thus establishes the electrochemical gradient responsible for driving ATP synthesis. Hdr splits apart the CoM-CoB heterodisulfide by reducing ferrodoxin and thus replenishes the ferrodoxin pool that is required to run the very first step of CO_2 reduction. This is in contrast to cytochrome-containing methanogens which are almost exclusively found within the *Methanosarcinales*. In these substrate-diverse methanogens, the Hdr enzyme evolved to harbor a cytochrome and can utilize methanophenazine as another electron carrier. Instead of directly reducing and replenishing the organism's supply of ferrodoxin, Hdr expels hydrogen ions to establish a proton-based electrochemical gradient that is used by a membrane-bound energy-conserving hydrogenase (Ech) to regenerate the reduced ferrodoxin. This system is best exemplified by the *M. barkeri* model (*i*AF692) in which the Ech reaction was of particular interest during model construction because the ratio of protons translocated to electrons extracted was unknown at the time. Using experimental growth yield data, a stoichiometry of 1 proton/$2e^-$ and GAM/NGAM of 70/1.75 mmol/gDWT/hr enabled the model to predict growth yields consistent with experimental data for growth on methanol, acetate, H_2/CO_2, and pyruvate. This Ech stoichiometry was later updated in *i*MG746 to 2

protons/2e$^-$. Although very closely related to *M. barkeri*, *M. acetivorans* has significant differences as a marine methanogen. Within methanogenesis, it substitutes Ech with the ferrodoxin:NAD^+ oxidoreductase complex (Rnf) which interestingly translocates sodium ions instead of hydrogen ions [296]. This establishes a primarily Na^+ dominated electrochemical gradient and helps explains why *M. acetivorans* inhabits a marine environment [195] in contrast to freshwater *M. barkeri* [297]. Since *M. acetivorans* is not able to consume CO_2, it would not be carrying out the endergonic first step of reducing CO_2 and thus justifies the absence of an Ech.

The majority of methanogenic GEMs available to date derive from the *M. barkeri* model *i*AF692 and the *M. acetivorans* model *i*MB745. Both models describe seven major metabolic subsystems: Vitamins and cofactor biosynthesis, Amino acid metabolism, Nucleotide metabolism, Central metabolism, Lipid and Cell wall biosynthesis, and Methanogensis. *i*MB745 inherited most of the reactions in *i*AF692 but also incorporated various additional pathways. The most notable changes include a modification of the methanofuran biosynthesis pathway based on homology of enzymes to those from the same pathway in *M. jannaschii*, a modified electron transport chain reflecting the aforementioned substitution of Rnf for Ech, and an updated biomass reaction that incorporated new carbohydrate, lipid, and nucleotide composition data. Although an attempt was made to estimate the GAM purely from genomic data, the model had to retain and optimize *i*AF692's original value in order to fit experimental growth data. The biomass reaction was more systematically constructed in *i*MB745 than *i*AF692. The general compo-

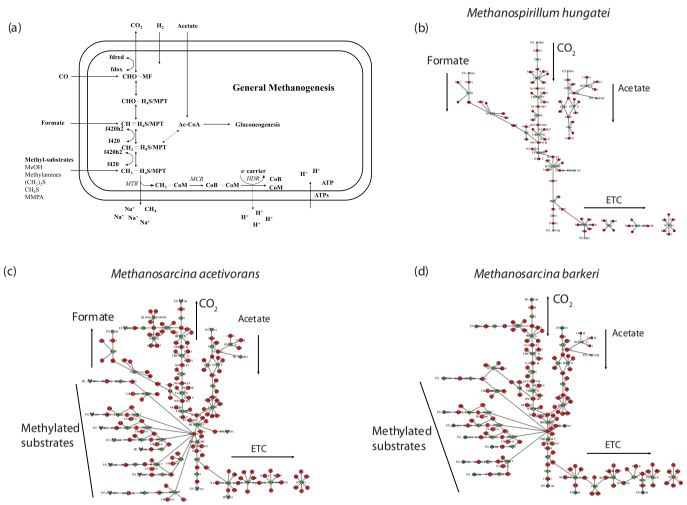


Figure 4.3: **Methanogenesis Pathway Framework.** (a) The basic structure of methanogenesis showing where the major growth substrates across all methanogens enter the pathway. The simplest methanogens are only capable of growing on CO_2/H_2 and/or formate while the most complex methanogens are also capable of methylotrophic and acetotrophic growth. H_4S/MPT stands for tetrahydrosarcinapterin (H_4SPT) and tetrahydromethanapterin (H_4MPT). The former is found exclusively in *Methanosarcinales* whereas the latter is found in all other methanogens. (b) The hypothesized methanogenesis pathways for *M. hungatei* (*i*Mhu428) [285]. Although it cannot use acetate as an energy source, the pathway to take up acetate is still present to shuttle it into gluconeogenesis. (c) The hypothesized methanogenesis pathway for *M. acetivorans* (*i*ST807) [35,36]. The conventional CO_2 reduction pathway is only run in reverse as this methanogen cannot metabolize CO_2. (d) The hypothesized methanogenesis pathway for *M. barkeri* (*i*AF692) [30] which bears great resemblance to that of *M. acetivorans*. The major differences between the two organisms' methanogenesis pathways lie in the electron transport chain (ETC). The specific pathways for each methanogen follows the same topological structure as the General Methanogenesis illustration. Red circles are metabolites while green diamonds signify enzymatic reactions of the pathway.

nents of the biomass reaction (Proteins, RNA, DNA, Lipids, Carbohydrates, and Trace components) were taken from a typical prokaryotic cell. Each biomass component required various metabolites within the network related to synthesizing that component. The protein component consisted of each individual amino acid and the biomass requirement for each was computed from its codon abundancy in the genome. The RNA and DNA biomass components were similarly computed from the nucleotide abundancy in the genome. The lipid biomass component consisted of experimentally characterized lipids within *M. barkeri* but the requirements of each lipid were adjusted to reflect their proportions within *M. acetivorans*. The carbohydrate biomass component consisted of four glycans and their requirements were set using the experimental compositions within *M. barkeri*. The trace component requirements of the biomass were directly taken from *i*AF692. Although the estimation methods to obtain the stoichiometric requirement for each individual metabolite in the biomass reaction is reasonable, it is important to note that the percentage of the general biomass components were adapted from an average bacterial cell instead of an average methanogenic archaea cell, most likely due to the lack of experimental data. This practice is quite common when reconstructing archaeal GEMs and can have serious consequences because the biomass composition has significant influence over metabolic flux distributions throughout the network. Without accurate experimental archaeal biomass compositions to guide the current models, other parameters in the models are tuned in order to fit flux predictions to various growth data. These fits may be biased by the use of bacterial biomass

compositions rather than archaeal biomass compositions (see Table 4.2 and 4.3 for biomass compositions from models).

*i*VS941 was developed and published independently from *i*MB745 at the same time through homology comparison with *M. barkeri* and an automated curation procedure published by Suthers, et al. [298]. The biomass reaction, which includes the GAM parameter, directly inherited from *i*AF692, but the nucleotide compositions were modified to reflect the differences in G/C content between *M. barkeri* and *M. acetivorans*. The most recent models of the *M. acetivorans* lineage are *i*MAC868 and our own *i*ST807, both of which are independently updated metabolic models. Although the *M. smithii* *i*Msi385 and *M. hungatei* *i*Mhu428 models are indeed independent curations, we will not discuss them here because they directly inherited from *i*MB745 and were qualified as preliminary draft models needing further revisions. *i*MAC868 was constructed to incorporate an engineered pathway that allowed for methane oxidation, essentially enabling the model to grow on methane and thus reverse the entire process of methanogenesis to produce the growth substrates that *M. acetivorans* would normally consume. Nevertheless, the model can still be used for simulations of a "wild-type" *M. acetivorans* and contains important updates to *i*MB745. *i*MAC868 merged the information from both *i*MB745 and *i*VS941 into a single model and corrected numerous charge and mass imbalances within the electron transport chain. 64 GPRs were also updated with the most recent *M. acetivorans* gene annotations. The biomass, GAM, and NGAM requirements remained the same as those

Table 4.2: **Cellular Molar Fractions.** The composition of each Archaea's biomass organized by major categories.

Molecule	*i*OG490 *i*OG478	*i*VS941	*i*MB745 *i*MAC868[a] *i*ST807	*i*AF692	*i*MG746	*i*Mhu428	*i*MM518	*i*Msi385	*i*OG654
Amino Acids	0.894	0.869	0.889	0.858	0.852	0.889	0.662	0.904	0.908
DNA	0.003	0.018	0.013	0.016	0.008	0.013	0.013	0.016	0.012
RNA	0.086	0.090	0.076	0.102	0.100	0.076	0.080	0.077	0.066
Lipids	0.015	0.009	0.008	0.009	0.009	0.008	0.277	0.002	0.013
Carbohydrates	NA	0.002	0.009	0.002	0.008	0.009	0.0004	NA	NA
Soluble Pool	0.001	0.013	0.006	0.013	0.016	0.006	0.0081	NA	0.001

[a]The *i*MAC868 model adopted the *i*MB745 biomass expression verbatim. [b]The *S. solfataricus* paper does not detail the biomass components and the model was not available to query. "Soluble Pool" includes various vitamins, cofactors, and trace metals.

Table 4.3: **Biomass Compositions and Energy Requirements** Molar composition of each Archaea's biomass for nucleic and amino acids. Additionally energy requirements for growth and persistence.

Molecule[a]	*i*OG490/*i*OG478	*i*VS941	*i*MB745/*i*MAC868[b]/*i*ST807	*i*AF692	*i*MG746	*i*JH428	*i*MM518	*i*Msi385	*i*OG654[c]	*i*TU515[d]
dAMP	0.0025	0.0360	0.0234	0.0331	0.0327	0.0232	0.0370	0.0331	0.027	NA
dCMP	0.0049	0.0234	0.0175	0.0215	0.0212	0.0176	0.0196	0.0215	0.016	NA
dGMP	0.0025	0.0234	0.0175	0.0215	0.0212	0.0176	0.0173	0.0215	0.016	NA
dTMP	0.0051	0.0360	0.0237	0.0331	0.0327	0.0236	0.0376	0.0231	0.027	NA
AMP	0.0618	0.0012	0.143	0.1846	0.1782	0.1152	0.1795	0.2222	0.155	NA
CMP	0.131	0.1637	0.115	0.1379	0.1361	0.1008	0.1780	0.1379	0.086	NA
GMP	0.131	0.2637	0.101	0.2222	0.2193	0.1200	0.1609	0.2222	0.086	NA
UMP	0.0619	0.1767	0.120	0.1489	0.1469	0.1440	0.1877	0.1489	0.155	NA
Ala	0.345±0.08	0.5621	0.388	0.5621	0.5546	0.3906	0.5853	0.5621	0.557±0.005	NA
Arg	0.211±0.044	0.3237	0.253	0.3237	0.3194	0.2520	0.2805	0.3237	0.378±0.022	NA
Asp	0.431±0.094	0.2638	0.301	0.2638	0.2603	0.3024	0.2805	0.2638	0.913±0.080	NA
Asn	0.099	0.2638	0.253	0.2638	0.2603	0.2520	0.2805	0.2638	NA	NA
Cys	0.033	0.1002	0.07	0.1002	0.0989	0.0693	0.0915	0.1002	0.056	NA
Glu	0.72±0.22	0.288	0.450	0.2880	0.2842	0.4473	0.3170	0.288	1.074±0.135	NA
Gln	0.125	0.288	0.143	0.2880	0.2842	0.1449	0.3170	0.288	NA	NA
Gly	0.29±0.053	0.6704	0.408	0.6704	0.6615	0.4095	0.4695	0.6704	0.561±0.049	NA
His	0.133±0.026	0.1037	0.094	0.1037	0.1023	0.0945	0.0976	0.25	0.104±0.009	NA
Ile	0.137±0.031	0.3179	0.415	0.3179	0.3137	0.4158	0.2683	0.3179	0.272±0.012	NA
Leu	0.251±0.052	0.493	0.534	0.4930	0.4865	0.5355	0.5182	0.493	0.471±0.029	NA
Lys	0.115±0.023	0.3755	0.370	0.3755	0.3705	0.3717	0.3292	0.3755	0.219±0.017	NA
Met	0.05	0.1682	0.132	0.1682	0.1660	0.1323	0.1341	0.1682	0.028±0.012	NA
Phe	0.111±0.022	0.2027	0.251	0.2027	0.2000	0.2520	0.1829	0.2027	0.242±0.018	NA
Pro	0.111±0.022	0.2419	0.225	0.2419	0.2387	0.2268	0.2073	0.2419	0.358±0.034	NA
Pyl	NA	NA	NA/NA/0.0808	NA	NA	NA	NA	NA	NA	NA
Ser	0.22±0.053	0.2361	0.390	0.2361	0.2330	0.3906	0.2683	0.2361	0.332±0.016	NA
Thr	0.181±0.036	0.2776	0.307	0.2776	0.2739	0.3087	0.2927	0.2776	0.415±0.011	NA
Trp	0.052	0.0622	0.060	0.0622	0.0614	0.0567	0.0061	0.0622	0.076	NA
Tyr	0.048±0.023	0.1509	0.210	0.1509	0.1489	0.2079	0.1585	0.1509	0.166±0.028	NA
Val	0.25±0.057	0.4631	0.387	0.4631	0.4570	0.3906	0.4085	0.4631	0.480±0.061	NA
GAM[e]	NA	70.0	65.0	70.0	65.0	47.0	29.8	50	30±4[c]	24.86
NGAM[f]	2.0	1.75	2.5	1.75	2.0	0.6	0.4	NA	2.0[c]	1.9

[a]Unless otherwise noted the units of the biomass coefficients are in units of mmol/gDCW. [b]The *i*MAC868 model adopted the *i*MB745 biomass expression verbatim. [c]The model was formulated in units of μmol/OD·L. [d]The *S. solfataricus* paper does not detail the biomass components and the model was not available to query. [e]Growth Associated Maintenance (GAM) has units of mmolATP/gDCW. [f]Non-growth Associated Maintenance (NGAM) has units of mmolATP/gDCW/hr.

from *i*MB745. In *i*ST807, we updated *i*MB745 by revising the methanofuran biosynthesis pathway with the most recent experimental data from *M. jannaschii* [201–204], adding 13 new reactions and 62 new genes, and revising the biomass reaction to utilize charged tRNAs instead of free amino acids. Among the new additions are reactions to enable pyrrolysine biosynthesis during methylamine growth, methyl-3-mercaptopropionate metabolism, and o-phosphoserine conversion to cysteine after aminoacylation [36]. Being able to uptake the various media components (Wolfe medium [178]) in which *M. acetivorans* is typically grown is crucial for accurately simulating the organism's metabolism. Many of the reactions required to emulate this are either missing or turned off in *i*MB745 and *i*MAC868. Cysteine is an important media component usually added with the purpose of quenching any oxygen in the methanogen's growth environment, but no one to date has verified whether this media component is also metabolized. Since unconstraining its uptake within *i*MB745 caused erroneously high growth rates, the cysteine uptake reaction was shut off and this was inherited by *i*MAC868 along with the various missing Wolfe media uptake reactions. *i*ST807 fixes this by incorporating uptake reactions for all the components of the Wolfe medium that have use in the metabolic network, including cysteine which is constrained to a non-growth-limiting value that maximizes the model's agreement with the experimental growth rates shown in Figure 4.4.

From the methanogenic GEMs geneology, it is clear that most of the methanogenic GEMs inherited from *i*MB745 despite the fact that *i*VS941

was independently published at the same time. This inheritance trend is most likely due to the more complete model documentations and availability of a readily testable GEM provided for *i*MB745 in contrast to *i*VS941. Given that metabolic modeling for archaea is still a developing effort, this practice of providing poorly assembled GEM files that are ill-prepared for quantitative assessment is still, unfortunately, common in the field. In order to alleviate this problem, we provide with this review all the currently available *M. acetivorans* models standardized to use BIGG IDs (`http://bigg.ucsd.edu/data_access`) and proper compartment tags such that the models can be conveniently handled within COBRApy. We also compare their growth characteristics as shown in Figure 4.4 to give a sense of how well these model perform with respect to each other and experimental data. We chose to focus on these models from this specie because they are often used as templates for the reconstruction of many other methanogens.

*i*VS941 predicts unrealistic growth rates, growth yields, and no methane efflux, which indicates the model's deficiency. This growth characteristic assessment shows that besides proper documentation, *i*MB745 also demonstrates better predictive ability over *i*VS941 and thus serves as a more reliable parent model for *M. acetivorans*. This is also evidenced by the predictive performances of *i*ST807 and *i*MAC868 which are updated versions of *i*MB745. Across all growth substrates and growth characteristics, *i*MAC868 predictions showed a median deviation of 36% from experimental values. In contrast, *i*ST807 demonstrated only a median deviation of 12% which is a

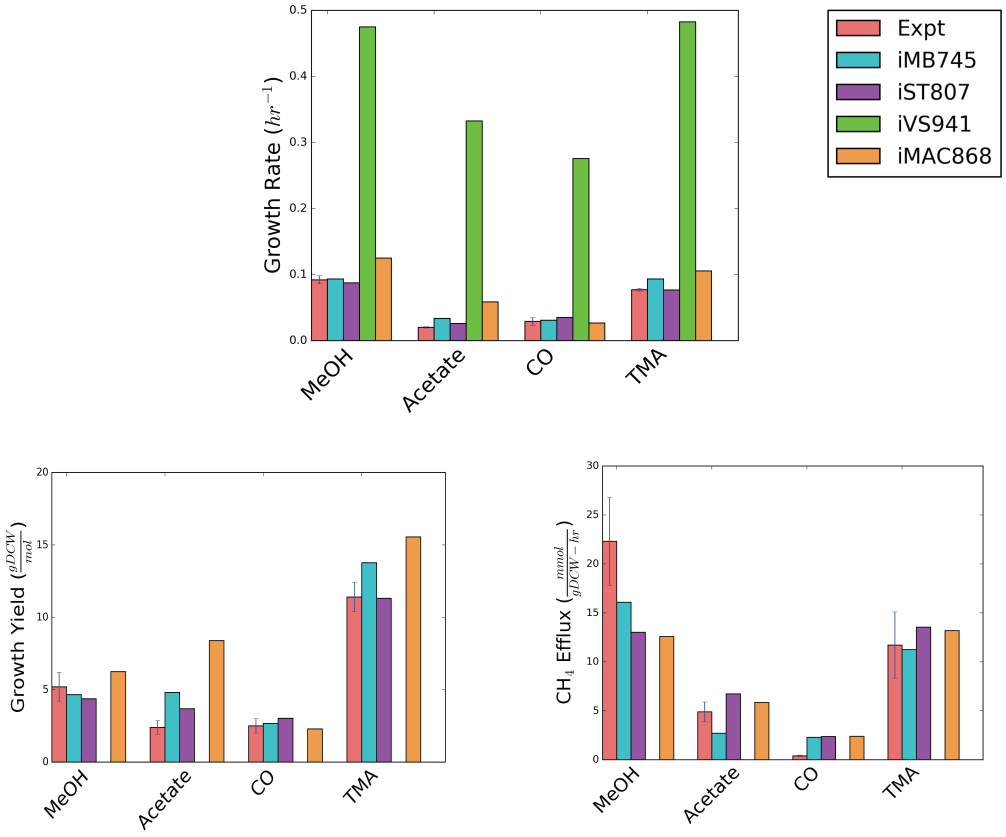


Figure 4.4: **Growth Characteristics of *M. acetivorans* Models.** The models were simulated using experimental growth substrate uptakes of MeOH:20, Acetate:7, and CO:11.6 mmol/gDCW/hr. Since experimental TMA uptake rates were not available, it was set to 6.77 mmol/gDCW/hr across all the models. This value was determined by fitting *i*ST807 to experimental growth rates on TMA. *i*VS941 gave unrealistically large growth yields and therefore the values were omitted from the Growth Yield plot for a clearer display of the other models' performances. *i*VS941 also did not predict any methane production under the given growth conditions. Experimental growth rates are from [17, 19, 23, 25, 104, 134, 143, 194, 195, 217, 235]. Experimental growth yields are from [195, 236, 237]. Experimental CH_4 production rates are from [104, 237–239]

marginal improvement over the 14% median deviation of *i*MB745. Although these statistics may seem to suggest *i*MB745 and *i*ST807 are more reliable models overall, it is important to keep in mind that growth predictions are heavily dependent on each model's allowed uptake reactions and their respective rates. In this assessment, each model's uptake reactions were set to the defaults that were provided within their respective publications. The uptake rate for the growth substrate being tested was uniformly set to the experimental value across all the models, and all other major growth substrate uptake reactions were turned off.

4.5 Non-Methanogen GEMs

4.5.1 *Halobacterium salinarum*

While only four GEMs have been developed for only three non-methanogenic archaea, they provided significant insight into the metabolism and growth of the the organisms. A reconstruction of the halophilic archaeum *Halobacterium salinarum* R-1 capable of growing on 15 different carbon/energy sources was developed by the group of Dieter Oesterhelt [290]. During reconstruction, a novel pentose phosphate pathway (PPP) for the generation of ribulose-5-phosphate (R5P) was predicted and later verified. It was known that different archaea used different pathways to produce R5P (e.g. non-oxidative PPP, reverse ribulose monophosphate pathway, and oxidative PPP) *H. salinarum* was missing all or portions of these pathways. An alternate pathway using

the partial Entner-Doudoroff (ED) pathway were connected to the partial oxidative branch of the PPP by a semiphosphorylated 6-phospho-gluconate. This pathway thus described why the organism retained parts of the oxidative PPP and part of the ED pathway even though it is incable of growing on sugars. During the reconstruction the authors also noted that shikimate production was incomplete and thus proposed that hexose and L-aspartate-4-semialdehyde were used, consistent with ^{13}C labeling data from tryptophan degradation. Additionally, draft pathways for synthesis leucine, isoleucine and valine could be generated in the model.

To calibrate the model they measured the amino acid composition and content using experiments and found protein mass constitute $\sim$49% of the dry mass, much less than in the other methanogens. Using dynamic simulations with experimenteally measured uptake rates for amino acids they predicted internal fluxes from which they drew a number of conclusions. Most strikingly, only 15% of amino acid cabons ended up in biomass with the majority being used to produce energy in the TCA cycle. They found that all amino acids were simultaneously used, though arginine, aspartate, leucine and isoleucine were taken up the most quickly, even the essential amino acids methionine, lysine, isoleucine, leucine and valine which the cells are incapable of producing. Using flux analysis they found that *H. salinarum* primarily produces isoprenoid lipids using leucine ($\sim$10%) while isoleucine was primarily degraded entering the TCA as acetyl-CoA and succinyl-CoA. Valine was the only amino acid that was primarily incorporated into biomass. Because the uptake rate of amino acid far outpaced the biomass incorpora-

tion they hypothesized that degradation pathways for all amino acids exist and proposed six enzymes to facilitate some reactions. However, it was only later that they determined the biosynthetic pathways for aromatic amino acids which they shared in common with *M. jannaschii*; during the discovery they used the metabolic model to identify uptake rates in auxotrophs [299]. Most impressively, they predicted, and later experimentally verified, that arginine is interconverted to ornithine during its degradation and is excreted to the environment early in growth, only to be taken up later as a source of arginine. Overall, they suggested that the greedy consumption of all available amino acids result in the "blooms" observed in the wild [290] and indicate that the metabolic pathways that have evolved are such that the organism can eat as quickly as possible to outgrow competitors.

The model was later updated to include a refined description of the respiratory chain as well as phototrophic growth leading to additional insights into metabolism [291]. Several key differences in the oxidative phosphorylation pathway compared to bacteria and mitochondria were proposed. First, that because complex I is missing the NADH oxidation subunits that it uses another energy carrier. Second, that halocyanin carries electrons from complex III to complex IV rather than menaquinone. Finally, that ATP synthase has a stoichiometry of 10 protons per ATP, which is much higher than in most organisms.

By fitting uptake rates of amino acids to aerobic growth experiment measurements, they identified isoleucine, leucine and valine as the preferred energy sources, while others such as alanine, proline and ornithine had dis-

tinct periods of different uptake rate [291]. Thus the organism hierarchically uses metabolites to maximize growth rate. They also predicted significant overflow of alanine, acetate and succinate. Interestingly, the identified that arginine fermentation essentially kick starts the cells growth, after which amino acid degradation and photosynthetic growth become dominant. They found that even during anaerobic phototrophic growth, the organism breaks down amino acids to obtain energy, even though they were incapable of deriving the maximal energy from respiration. Interestingly, they could identify that the network structure of amino acid degradation could describe why alanine was produced; specifically, as an overflow pathway during serine consumption. This is in contrast to aerobic growth where serine and alanine consumption appear to coincide with one another, likely due to the fact that pyruvate can be pushed funnelled into the TCA cycle. Overall, the studies of *H. salinarum* led to the conclusion that the organism evolved its metabolic behavior to maximize growth during blooms, which can occur sporadically with many years in between [290, 291]. It was suggested that they use this as a strategy to out-compete other organisms that feed on available nutrients and build up enough of a population that they can survive long periods of starvation [291].

4.5.2 *Natronomonas pharaonis*

The metabolic network for the polyextremophile (high salt concentration and alkaline pH)*Natronomonas pharaonis* was developed [294] using the reconstruction for *H. salinarum*. The network is significantly larger, with nearly 30%

more genes associated with reactions, mostly due to additional amino acid and carbon degradation pathways. As *N. pharaonis* is capable of growth on a single carbon source the reconstruction complements that for *H. salinarum* which requires a complex broth for growth. For this reason, the reconstructions could be used to investigate questions regarding the metabolic objective of halophiles that are subject to different evolutionary pressures and answer questions about optimality of energy production.

The authors measured the amino acid content to define the biomass composition and found, similarly to for *H. salinarum* that it made up about 75% of organic mass [290,294]. Perhaps the high protein content helps to compensate for the high osmolarity in which the organisms are grown. Using the model predictions about aerobic growth were obtained, most importantly that at very high (>7:3) and low (<3:7) acetate to oxygen consumption ratios the organism was incapable of growth. Using experiments, they identified an acetate:oxygen ratio of about 1:2 and an ATP maintenance cost of $\sim$30 $\mu mol/\Delta OD \cdot ml$. A wide range of maintenance energies and acetate:oxygen ratios gave near optimal growth, indicating that growth of *N. pharaonis* is robust to environment and the biological objective is maximizing growth and energy production [294]. They found that the carbon incorporation was actually quite low ($\sim$35%). Finally, using arguments about respiratory exchange ratio (e.g. the ratio between CO_2 production and oxygen consumption) the authors were able to demonstrate that about 10% of carbon is neither incorporated in biomass nor respired, suggesting that the organism uses some form of overflow metabolism [294]. While they did not make

any suggestions, there are a number of likely suspects such as succinate or pyruvate which could act as available nutrients for other organisms.

4.5.3 *Sulfolobus solfataricus*

The final nonmethanogen model developed for an archaea is for the hyperthermoacidophile *Sulfolobus solfataricus* [295]. The model and organism are remarkable among the archaea represented here in that they grow optimally at a pH of 3.5 and temperature of 80° and consume 35 different carbon sources. The thermostabilitiy of their enzymes are of interest to bioengineers and makes the organism attractive for bioreactor design. Their unique abilities give them an edge in the hot-springs where they are found and allow them to consume a plethora of degraded organic mass. The final reconstruction consist of 706 reactions associated with 515 genes and conveys the ability to consumer all 35 carbon sources. The model was calibrated with growth and non-growth associated maintenances of 24.68 and 1.9 mmolATP/gDCW/hr respectively to match experiments. Interestingly, the GAM is the largest of any of the archaea, while the NGAM is the smallest. Unfortunately, the model itself was not available and thus the biomass composition used in the study could not be compared with the others to identify the source of this low cost for growth (see Table 4.2). It could be due to the fact that the genome was relatively small (2.9 MB) compared with many of the other organisms. The authors of the study chose a phosphate/oxygen ratio of 0.5 as the final fit parameter of their model; this low value was due to the fact that the archaea uses inefficient cytochrome complexes SoxABCD and

SoxEFGHIM for respiration. Using these parameters, the model incorporates about 25% of carbon while respiring the rest.

During the model reconstruction, the authors identified the fact that *S. solfataricus* uses a reverse ribulose-monophosphate pathway (RRMP) instead of the pentose phosphate pathway. Specifically, they found that the organism was missing a transaldolase and thus they allowed accumulation of sedoheptulose 7-phosphate. They found that accumulation of sedoheptulose 7-phosphate in their simulated media accounted for $\sim$3% of all carbon atoms, and thus is a significant portion of the overall carbon available for biomass. Simulations indicated that on glucose growth about 22% of carbon flux was fed into the RRMP pathway while the rest was metabolised to pyruvate via the Entner-Doudoroff (ED) pathway to be subsequently used in TCA cycle. Flux variability analysis of the metabolic model demonstrated that both the semi-pphosphorylative and non-phosphorylative branches of the ED pathway were possible and indicates that further studies are required to understand the growth of the organism. Similarly, the TCA cycle showed significant variability, primarily due to the glycoxylate shunt. Finally, variability in the production of amino acids such as histidine, tryptophan, alanine and glutamate indicate different routes of synthesis.

Because a related organism *Sulfolobus sp.* VE 6 could grow autotrophically fixing bicarbonate, the authors searched for the hydroxypropionate-hydroxybutyrate cycle. They found 11 of the 16 enzymes and performed BLAS searches to identify putative homologs of the 5 remaining enzymes. Thus they predicted that *S. solfataricus* is able to grow autotrophically and

suggested experiments should be performed. During autotrophic growth it was predicted that the TCA cycle was little used with flux flowing from succinyl-CoA through malate to form pyruvate which could be used in gluconeogenesis. Additionally hydrogen sulfide was fixed to provide a sulfur source, and in fact produces energy allowing the simulated organism to grow much more quickly than on glucose; however, this is likely due to the lack of an uptake rate on H_2S. Regardless, this hints the possibility of syntrophic interactions with sulfate reducing bacteria.

The authors went on to compare the growth of the organisms on the 35 different carbon sources. To do this they fixed the carbon uptake rate and compared biomass flux. Overall, the organism grew significantly faster when growing on glycerol and propanol and marginally better on oligosaccharides. They also grew significantly more slowly on carbon sources that on compounds that enter the TCA cycle at points other than 2-oxoglutarate. Gene deletion assays indicated that over 50% of all single gene knockouts were nonlethal and an additional $\sim$25% allowed limited growth, suggesting *S. solfataricus* is metabolically versatile, a trait of potential use in high temperature environments where enzyme efficiency could be significantly lower and mutation rates could be much higher. Overall, these results suggest that *S. solfataricus* likely preferentially consumes certain carbon sources and likely regulates alternative pathways, thus leaving room for other niche organisms to grow in concert with them.

4.6 Comparison of Metabolic Capabilities

Well-curated metabolic models function as comprehensive databases of the knowledge about organisms; thus, they are potentially useful tools for studying evolution and diversity of metabolism. Three properties of the metabolic models are of particular utility for comparative studies: 1) metabolic models connect gene function with metabolic function via their gene-protein-reaction rules, 2) the metabolic network is topologically defined by metabolites and the reactions that convert them, rather than the genes that facilitate those interconversions, and 3) metabolic networks coupled with modeling techniques allow for the identification of function redundancy/degeneracy. The first of these properties allow direct comparison of gene content based on metabolic function and the application of traditional evolutionary tools (e.g. bioinformatic and phylogenetic approaches). The second of these properties allows networks to be compared by function rather than gene content; for example, the network could be used to identify convergent evolution. The final of these properties could be used to provide insight into the selective pressures of the organism; specifically, duplicated functionality might suggest a critically important function for the organism.

To demonstrate the utility of metabolic models to evolutionary analyses we computed the conservation of genes facilitating metabolic reactions. An ITEP database [191] including each of the organisms in Table 4.1 was constructed using the default parameters. Briefly, ITEP is a software toolkit for examining microbial pan-genomes that provides functionality

for constructing a BLAST database and querying protein family prediction, ortholog detection, and analysis of functional domains. Among its capabilities is assessing the GPRs of a metabolic model for each reaction and determine whether or not the homologs exist in another organism. The GPRs from the *M. acetivorans* model *i*ST807 were used as input to the `db_evaluateReactionsFromGpr.py` function to assess the conservation in other organisms. The "or" option to the function was used to assess whether any genes for each *M. acetivorans* reaction existed in the other archaea discussed in this review. Doing this for each organism, we computed the extent of conservation for each reaction (e.g. the fraction of organisms in which the reaction had conserved genes). The results can be seen in Fig. 4.5. An examination of the figure shows that most reactions were either highly conserved (red) or very lowly conserved (blue) with few that were conserved among some of the species. Highly conserved reactions appeared in central amino acid biosynthesis, nucleotide metabolism, tRNA charging and frustose metabolism. Conversely, those in transport, specialized lipid metabolism and vitamin and cofactor metabolism (especially methanogen specific pathways such as coenzyme F420, coenzyme F390 and adenosylcobalamin biosynthesis) were not highly conserved.

Categorizing the homologous genes computed by ITEP by metabolic subsystem—as annotated in *i*ST807—lends more specific insight into conservation of metabolism in these archaea (see Fig. 4.6). Amino acid biosyntehtic pathways are generally highly conserved (labelled in blue if Fig. 4.6). Proline biosynthesis is a notable exception; none of the genes annotated in *M.*

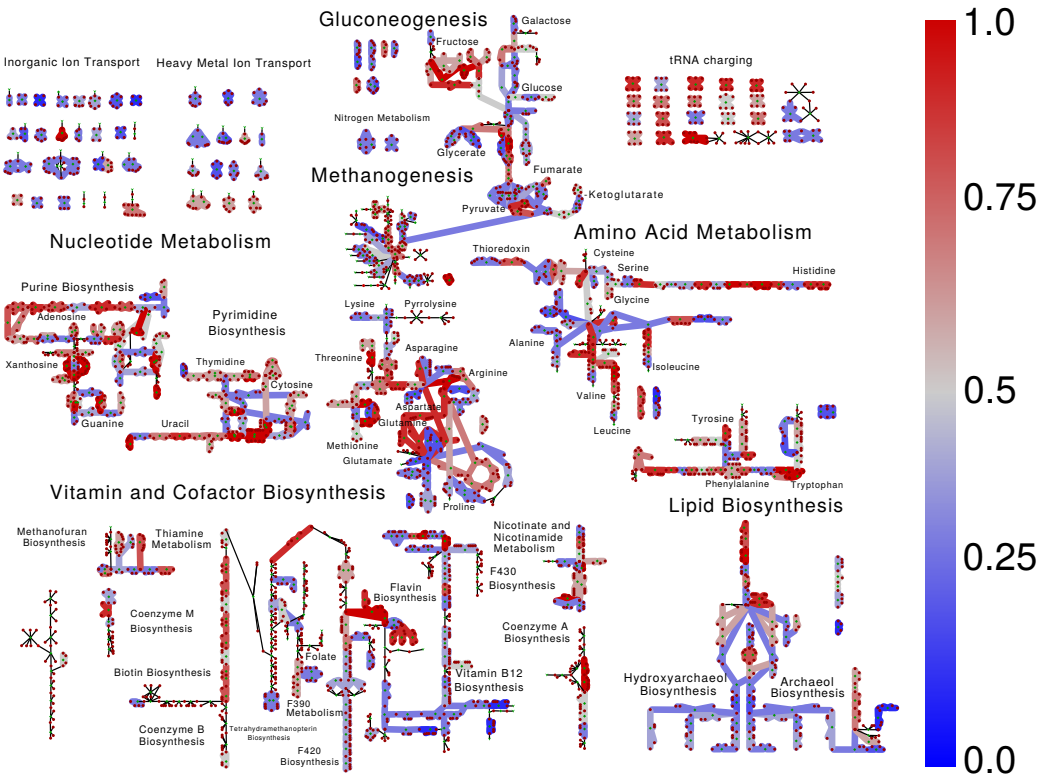


Figure 4.5: **Conservation of Metabolic Reactions.** A map showing the extent of conservation for the reactions of the *M. acetivorans* model *i*ST807 (as encoded in the gene-protein-reaction associations (GPRs) of the model). Nodes represent either a metabolite or reaction and edges indicate the dependencies between reactions and metabolites. Reactions on the blue end of the spectrum are facilitated by enzymes that are conserved in relatively few of the archaeal species studied, while reaction in red are facilitated by highly conserved enzymes. Reactions with thin grey lines are not associated with genes. To assess conservation, the `db_evaluateReactionsFromGpr.py` functionality of the ITEP software [191] was used. It computes homologous genes to those in the GPRs of each reaction in *i*ST807. The ITEP function was executed with the "or" option enabled to identify whether *any* of the enzymes (or enzymatic subunits) annotated as facilitating the reactions were encoded in the organism.

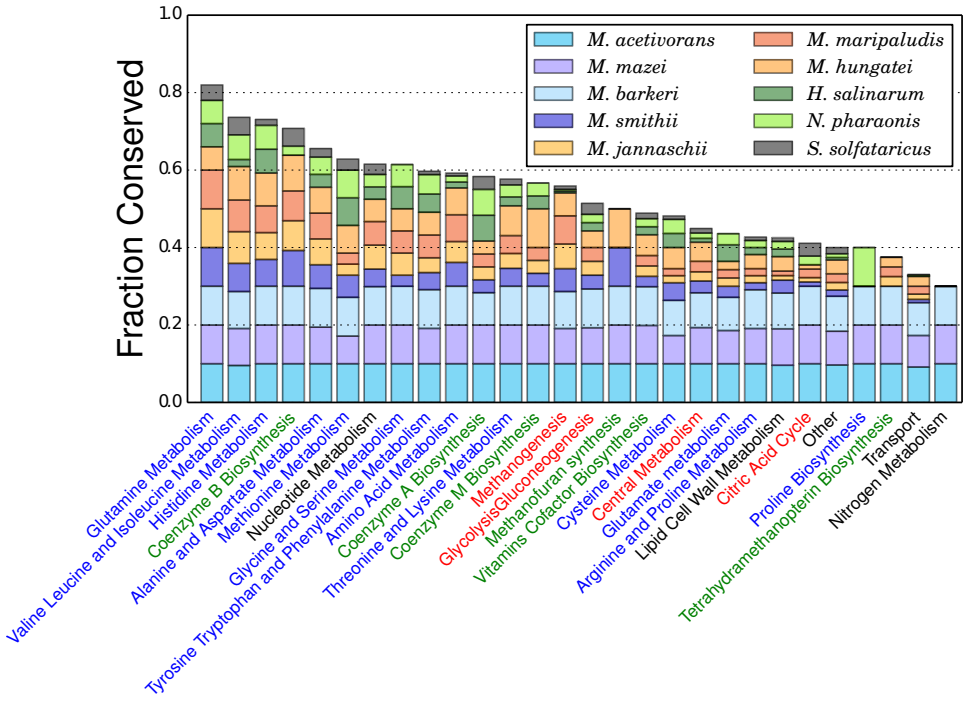
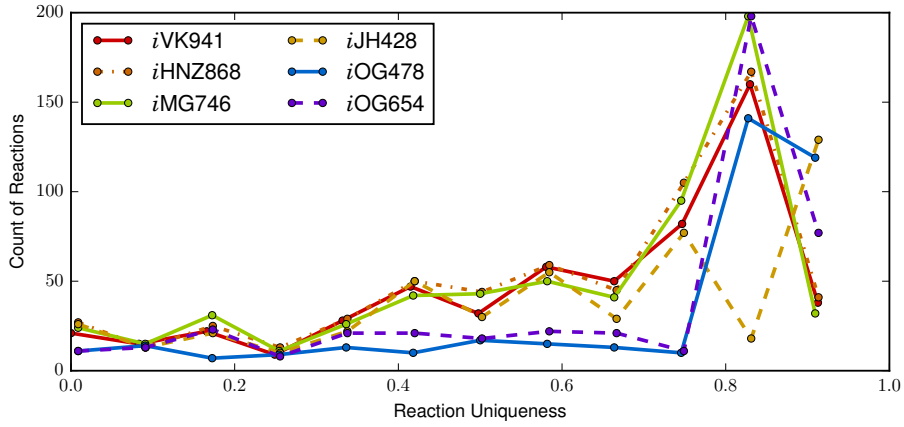


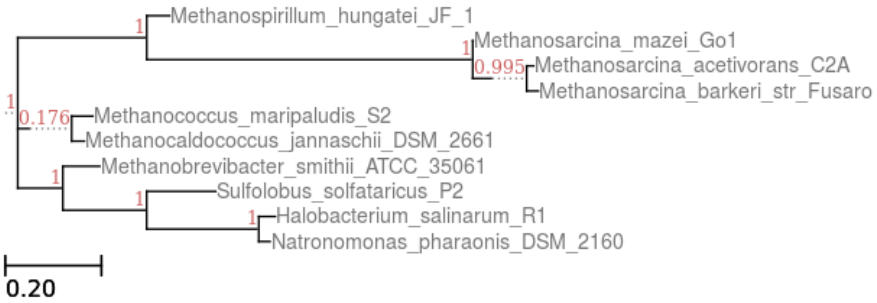
Figure 4.6: **Fraction of Conserved Reactions.** Grouping of the information shown in Fig. 4.5 by metabolic subsystem (as annotated in *i*ST807). Type I and type II methanogens are shown in warm and cool colors, respectively, while green show halophilic archaea. The overall height of each bar indicates the total fraction of reactions in the metabolic subsystem that are conserved, while the height of individual portions of each bar indicate the relative conservation of reaction in the subsystem from that organism. Metabolic subsystems labels are color coded: amino acid metabolism in blue, vitamin and cofactor metabolism in green, central metabolism in red and other categories in black.

acetivorans were found in type I methanogens, *H. salinarum* or *S. solfataricus*. Additionally, all genes annotated as synthesizing serine or glycine are missing *S. solfataricus*. This could indicate either an incorrect annotation in the model or multiple proline biosynthetic pathways in metabolism. Biosynthesis of methanofuran, a cofactor in methanogenesis, shows an interesting pattern of conservation: it is conserved in type II methanogens and *M. hungatei* but not the other type I methanogens. Notably nitrogen metabolism is the least similar among the Archaea as other have previously identified [300]. While these are broad statements, they demonstrate how using information from metabolic reconstructions can be quickly used to compare differences between metabolism and study evolution and conservation of metabolic pathways.

A similar analysis was performed with the GPRs from each of the metabolic models. The aggregate statistics from this analysis can be seen in Fig. 4.7a where the count of reactions with a particular conservation level are shown. Similarly to what was seen for *M. acetivorans* there appear to be mostly reactions that are highly conserved (low reaction uniqueness) or very lowly conserved (high reaction uniqueness). This is not a surprising observation as it has long been known that metabolic networks have a bowtie topology [301] and are generally scale-free networks [302]. Cross comparison between each of these conservation predictions in fine detail is beyond the scope of this review; however, we feel we have demonstrated the utility of using metabolic reconstructions as a tool to compare metabolism.



(a)



(b)

Figure 4.7: **Diversity and Phylogeny of Metabolic Models.** (a) The uniqueness of the reactions in each of the metabolic models. Here, we define uniqueness to be the fraction of archaea (227 in total with closed or nearly complete genomes) that do not have the reaction; higher uniqueness means fewer of the organisms contain the genes coding enzymes that are annotated by the GPRs of the specified models. The presence of the genes are computed using ITEP as discussed in Figure 4.5. (b) A phylogenetic tree computed based on similarity to the *M. acetivorans* model *i*ST807. This tree is based on the ITEP results discussed in Figure 4.5.

4.7 Conclusions

We have presented an overview of genome-scale metabolic models and discussed the defining metabolic features among the few GEMs available for archaea. In these discussions, we have also highlighted some much-needed improvements to model building practices in order to facilitate the development of archaeal models. By using the gene-protein-reaction associations in these archaeal GEMs, we also demonstrate the invaluable utility of these metabolic models as they can be extended beyond flux analysis to gain significant insight into evolutionary patterns among organisms. Visualizing the known archaeal metabolic models on a phylogentic tree (see Figure 4.1) leads to the conclusion that model development in the community thus far has mostly focused on Euryarchaeota, leaving the Crenarchaeota largely unexplored. Although models do not yet exist for members of the Archaeoglobi, Thermoplasmata and Thermococci classes, all other major Euryarchaeota classes have at least one representative model. This is not to underestimate the importance of further developing these Euryarchaeota models as Archaeoglobi have some of the most diverse metabolisms of any Euryarchaeota, capable of chemolithotrophy by reduction of sulfates, thisulfates, nitrates and heterotrophy via reduction of sulfates via organic compounds [250]. However, the paucity of GEMs for the Crenarchaeota is a major impediment for a comprehensive study of evolution and diversity in Archaea. GEMs are invaluable tools to help guide the exploration and comparison of the great metabolic diversity of energy conservation in these

organisms which are capable of sulfate reduction both chemolithotrophically and heterotrophically (members of the Desulfococcales), nitrate reduction (members of Thermoproteales) hydrogen oxidation and sulfur reduction (members of the sulfolobales) [250]. The existence of diverse energy conservation pathways will likely come with diverse electron transport chain and transport systems. Understanding these unique characteristics will be paramount in understanding growth in extreme conditions and syntrophy among microorganisms as well as for engineering communities for biotechnology applications.

Chapter 5

A Pan-Genomic Comparison of the *Methanosarcina* Genus Through the Lens of Genome Scale Metabolic Modeling

Species of the genus *Methanosarcina* are the most metabolically diverse methanogens; they containing all four methanogenic pathways, are capable of growing on numerous substrates, and possess many of the largest archaeal genomes sequenced to date. As such, they exist in a variety of environments around the world. Yet much remains unknown about their metabolisms far to the periphery of the methanogenesis pathways. To expand our knowledge of the capabilities of the *Methanosarcina*, a pan-genomic examination of 30 genomes—27 of which are newly sequenced for this work—is performed, with specific focus on metabolic functions. A core-genome consisting of 1329 genes conserved among all the species is identified. The core is dwarfed by a variable pan-genome comprising five times as many genes. Five major groups, which cluster by metabolic capabilities, arise from this analysis, typified by the *M. barkeri*, *M. calensis*, *M. mazei*, *M. si-*

Work includes material and contributions from Matthew N. Benedict, James R. Henriksen, Judy Luke, Mary-Beth Metcalf, Sarah Stevens, Nicholas Youngblut, Nathan D. Price, Zaida Luthey-Schulten, Rachel D. Whitaker, and William W. Metcalf. Specifically, J.R.H, J.L., M.B.M., S.S., N.Y. helped with sequencing and construction of genomes, and M.N.B. worked closely in the analysis and construction of the metabolic models. M.N.B. also created figure 6.

ciliae, and *M. thermophila*. Genome scale metabolic models of each species are generated by propagating pre-existing methanogen models followed by manual curation. Analysis of the resulting models revealed key differences among the *Methanosarcina* spp. and, importantly, identified key gaps in metabolic knowledge that were inconsistent with experimental observations. Further, during the model construction process several conserved metabolic pathways absent in prior reconstructions were identified and added; most notably, for molybdopterin biosynthesis. Finally, by examining conserved genes, predictions about novel gene functions could be made, with nearly 150 new gene associations/functions hypothesized. Among them, is a gene encoding an enzyme that we predict catalyzes the final step of the methanofuran cofactor biosynthesis. Additionally, a pathway for the biosynthesis of methanophenazine–the final uncharacterized methanogenesis cofactor–is proposed based on conserved genes.

5.1 Introduction

Genome-scale metabolic modeling (GSMM) is a powerful way to consolidate large quantities of biological information in a way that enables phenotype predictions [303]. The resultant metabolic models have applications ranging from biotechnology to medicine [304–306], with new applications for the approach being developed every year [307, 308]. The utility of metabolic models has created a demand for methods that allow rapid extrapolation of models from well-studied organisms to their less well-studied relatives.

Although many metabolic enzymes are conserved across large swaths of the tree of life, there is significant variability in metabolic pathways, even among closely-related organisms [309]. Comparative genomics (CG) can be used to predict the extent to which metabolic reactions in one metabolic network are present in related organisms. Metabolic networks built based on comparative genomics approaches can capture much of the pathway variability present in these organisms, as determined by building and simulating constraint-based models based on them [310,311]. The advent of low-cost, high-throughput DNA sequencing, the continued development of manually-curated metabolic networks [312], and the advent of automated network reconstruction and model building strategies [273,313,314] have made it increasingly feasible to use this approach to study phenotypic variation at a whole-clade scale.

Although comparative modeling is able to accurately capture many phenotype differences between related organisms, the accuracy of propagated models depends on the quality of both the reference networks and the quality of genome sequences that are used to propagate them. The quality of genomic data can vary widely because of limited experimental resources, technical limitations of different sequencing platforms, and divergent strengths and weakness of various gene calling and annotation algorithms [315]. Metabolic network quality and genome coverage also varies widely between reconstructions, depending on the depth of biochemical data available to support them [308]. As part of the curation process for metabolic networks, it is necessary to distinguish between experimental or modeling artefacts and real differences that lead to phenotypic insight.

Comparative genomics techniques leveraging improvements in genome sequenceing technology to improve annotations, assemblies, and gene calls [316–318] are instrumental for making such distinctions. Because an organism's metabolic capabilities are correlated with its genomic content and because its genomic content is correlated with its phylogeny, combining comparative genomics of many related organisms with modeling can further enhance the ability of comparative genomics to separate artefacts from real metabolic divergence [319].

The genus *Methanosarcina* is an ideal test case for an in-depth assessment of comparative modeling's utility. Manually-curated, genome-scale metabolic models have been built [30, 31] and recently updated [36, 284, 286, 292] for two organisms in the genus: *M. acetivorans* C2A and *M. barkeri* str. Fusaro. These models capture important metabolic differences between these organisms that lead to clear phenotypic differences. For example, known differences in the energy conservation and enzyme activity between the two species explain the inability of *M. acetivorans* C2A to utilize to utilize hydrogen as an electron donor for methanogenesis [20, 320]. Although methane production and energy conservation has been well studied in these organisms [321], other metabolic pathways have not been adequately characterized. Hence, an opportunity exists to identify cases of diverged pathways and incorrect or incomplete metabolic gene predictions based on homology with bacterial genes.

In this study, we demonstrate the use of comparative genomics to identify problems in genome-scale metabolic models, leading to increased confidence

in the metabolic reconstruction and to significantly expand our knowledge of peripheral metabolic pathways. To do this, we have sequenced the genomes of 27 *Methanosarcina* species covering the diversity in the genus. The genomic data includes high-quality sequences (20 of which are closed) and 7 draft-quality genomes. We built metabolic models for each genome by propagating the metabolic networks for *M. barkeri* str. Fusaro and *M. acetivorans* C2A as references [36, 286, 292]. Doing so, we define the core- and pan-reactome for the *Methanosarcina* species. By comparing the phenotypic predictions of the constructed models with known physiology, we show that numerous predicted differences in metabolism between species are incorrect due to: 1) incomplete knowledge of archaeal metabolism, 2) incorrect assumptions in the reference models, or 3) problems with the genomic data. Moving beyond modeling known pathways, we analyzed the core- and pan-genome for species-specific differences in metabolism, which resulted in significant expansion of the metabolic models. Several missing pathways were identified and added, in addition to with a plethora of previously uncharacterised gene associations and reactions. Further, an analysis of the core-genome led to hypotheses about which genes catalyze key metabolic steps in, for example, the methanofuran and methanophenazine biosynthesis pathways.

5.2 Materials and Methods

Briefly, the genomes of three previously-sequenced *Methanosarcina* species were downloaded from GenBank [322]. Genome sequencing for 27 addi-

tional *Methanosarcina* species was performed as described in the following section and deposited in GenBank (accession numbers can be found in S1 Appendix Table 5.1). The genomes were annotated using RAST [323] and post-processed to remove very short hypothetical proteins (<200 BP), remove completely overlapping genes, and fix calling of pyrrolysine-encoding genes. Orthologous gene families in the *Methanosarcina* and greater Methanosarcinales were predicted using OrthoMCL [324] and analyzed using the ITEP toolkit [191]. Models of each sequenced species were built by propagating orthologous proteins from previously-published models of *M. acetivorans* C2A and *M. barkeri* str. Fusaro [284,292]; Manual curation of conserved genes and models drastically expanded modelled capabilities. Phenotype simulations were performed using Flux Balance Analysis [69] as implemented in the COBRApy [230]. Detailed procedures, comparative genomics methods, and construction of COBRA models used in the analysis are described in detail in the following sections.

5.2.1 Genome Sequencing, Assembly and Annotation

Previously Published Genomes

The genomes for *Methanosarcina acetivorans* C2A, *M. barkeri* str. Fusaro, and *M. mazei* Gö1 have been previously sequenced and published [14,321,325]. These genomes were downloaded from Genbank [322] and used without modification.

Methanosarcina **Typestrains**

The 27 *Methanosarcina* typestrains whose genomes were sequenced in this project had all been previously isolated [195, 297, 326–347] and were ordered from DSM). Each of these strains was grown on methanogen medium with methanol as a carbon source.

The extracted DNA was subjected to 454 Pyrosequencing (Roche) to $>30\times$ coverage using half shotgun and half paired-end reads. Draft genome assemblies were built from 454 reads using Newbler v2.3 (Roche).

Cosmid libraries were prepared for each genome. The ends of these cosmids were sequenced and used to scaffold the draft assemblies. To fill gaps in the draft assemblies, two technologies were used. When the gap was spanned by a cosmid, the spanning cosmids were subjected to sequencing with MiSeq (Illumina). Gaps that were not spanned by any cosmid or which failed to close using MiSeq sequences were closed using Sanger sequencing. Gap filling assemblies and draft contigs were manually merged using the CLC Genomics Workbench v3.3 (CLCBio) and Geneious v4.5 (Biomatters Ltd.).

Illumina data (HiSeq2000) was also generated for each genome. This data was used with the iCORN program [348] to correct assembly and sequencing errors.

5.2.2 Genome Annotation

Gene calls and annotations were performed using the RAST server [323] with FigFAMs v59. Two post-processing steps were performed on the called genes. First, all proteins annotated as 'hypothetical protein' with a length less than 200 BP and genes that completely overlapped with other called genes were removed (this step is consistent with that done for the *M. acetivorans* C2A genome [14]). Secondly, pyrrolysine-containing proteins [349] were fixed to correctly include the amber codon.

5.2.3 Prediction and Analysis of Orthologous Groups

Sequence similarity between each pair of genes was computed using BLAST+ 2.2.28 [350, 351] with an E-value cutoff of 10^{-5} and otherwise default parameters. Orthologous groups were predicted from BLASTP results using OrthoMCL 2.0.9 [352] coupled to MCL version 11-294 [353,354]. The default MCL inflation value of 1.5 and percent match cutoff of 50% were used for OrthoMCL computations. Orthologous groups were analyzed using the ITEP toolkit [191].

A concatenation of the amino acid sequences of 1329 completely conserved genes (see S1 Table) was aligned using Mafft v7.123b with default parameters [355]. PhyML v20131022 [356] was used to compute a phylogenetic tree from the alignment. The approximate likelihood ratio test (aLRT statistics) [357] was computed and used as a measure of support for the resulting tree.

Orthology predictions for selected metabolic genes were verified by computing the phylogeny of the broader protein families and correcting the predictions if needed. Broader families were identified by running BLASTp between members of an OrthoMCL-predicted orthologous group and the *Methanosarcina* proteomes. After curating orthology predictions, each putatively absent metabolic gene was verified to be absent from the assembly using tBLASTn [358].

5.2.4 Model Propagation

Models were generated in an iterative fashion shown schematically in Fig 5.1. Previously-published *M. acetivorans* C2A and *M. barkeri* str. Fusaro genome-scale metabolic models (*i*MG746, *i*MAC868, *i*ST807, *i*MG746) [36,284,286,292] were downloaded from the publications and adjusted to ensure consistency in metabolite and reaction nomenclature. Subsequently, lists of metabolic reactions in these organisms and their curated gene associations were compiled. Gene associations in the models are represented in the form of gene-protein-reaction relationships (GPRs), which are Boolean expressions describing the relationships between genes and reactions [359]. Genes that must all be present in an organism to perform a given reaction are given an AND relationship, while sets of genes that are each sufficient to perform the reaction are given an OR relationship.

OrthoMCL clustering results were used to evaluate whether an ortholog for each gene in each GPR was present or absent in each organism. If an ortholog was predicted to be present, the gene was given a Boolean value of

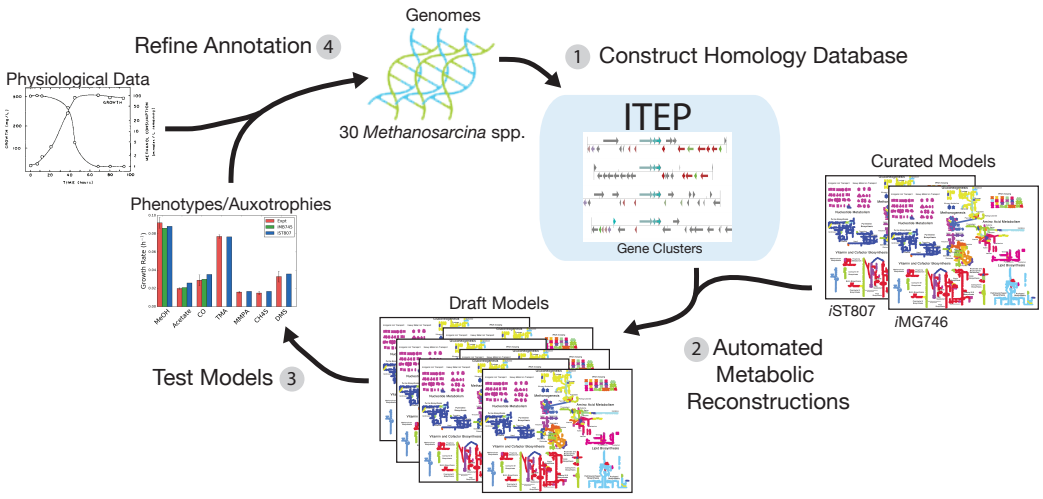


Figure 5.1: **Schematic of Model Construction and Refinement** A model-driven approach for generating/refining genome scale metabolic models. First, annotations for genomes from 27 *Methanosarcina* species were used to generate a homology (ITEP) database [191]. Next, highly curated metabolic models for *M. acetivorans* C2A [36, 284, 286] and *M. barkeri* str. Fusaro [292] were used as inputs to ITEP to generate draft models for the other 27 methanogens. Draft models were tested for correct prediction of growth phenotypes (*i.e.*, auxotrophies, growth substrates, *etc.*). Incorrect predictions directed targeted reannotation missing or incorrect predictions in draft genomes. Finally, manual curation was applied to identify previously uncharacterised reactions/metabolic capabilities and gene associations.

TRUE and if it was predicted to be absent, it was given a Boolean value of FALSE. The Boolean GPR expression was evaluated using these truth values to determine presence and absence of each reaction.

A model was built for each *Methanosarcina* spp, consisting of: 1) all reactions whose GPR evaluated to TRUE (present), and 2) all non-gene associated reactions from the *M. barkeri* str. Fusaro and *M. acetivorans* C2A models. Non-gene associated reactions include metabolite exchange reactions, gap-fill reactions and the biomass reactions. Many of these non-gene associated reactions have non-genetic evidence for their existence in other organisms [284,360] and produce essential cofactors such as coenzyme F_{420}.

The other *Methanosarcina* models were given the same reaction bounds and ATP maintenance parameters as the published models, except that the nonessential coenzyme F_{420}-dependent sulfite reductase (SULR2) was given a very low reaction rate (0.01 mmol/gDW/hr) and coenzyme F_{420} dehydrogenase was limited to a rate of -1 mmol/gDW/hr in the less favorable direction. Both of these changes were made to avoid ATP generating cycles in the new draft models.

5.2.5 Manual Curation

Phenotypic and growth simulations were performed on genome-scale metabolic models using Flux Balance Analysis (FBA) [69]. FBA Simulations were performed using GLPK v4.39 in COBRAPy [230]. The minimal media composition was as previously described [36], excluding non-essential growth factors. Simulations were performed using methanol, methylamines, dimethylsul-

fide, methanethiol, methylmercaptopropionate, acetate, and H_2/CO_2 as substrates. A list of reactions essential for growth was obtained by iteratively setting the maximum and minimum rates of each reaction to 0 and trying to maximize production of biomass on each of these substrates. The *M. acetivorans* C2A biomass equation from *i*ST807 [36] was used as the growth objective. A maximum biomass production rate of less than 10^{-5} hr^{-1} was considered lethal. Essentiality of each biomass component that a model could not create was evaluated manually based on known biological capabilities, and biomass equations (see Results Section 5.3.5).

Potential metabolic reactions that correct the known discrepancies between predicted growth and experiments were identified using the gap-filling algorithm implemented in COBRApy (which is based on the methods of Reed *et al.* [361] and Kumar *et al.* [362]). Gap-fill reactions were then used in a targeted search to identify potential genes associated with the missing functionality. A final manual curation step was performed by identifying all the of completely conserved genes within each metabolic group which had clusters of orthologous groups (COG) codes [196] associated with metabolism (see S2 Table). Those genes with sufficient similarity (*i.e.*, specific COG codes, protein families, or homology, *etc.*) to known metabolic genes in other organisms were assigned functions within the model either by assigning them to existing reactions or adding new reactions as needed. Overall, new characterizations of gene functions significantly expanded our knowledge about methanogen metabolism with 120 newly added genes due to new reactions, 90 new genes in gene-reaction associations, and 52 new

metabolites added in aggregate over the five groups of *Methanosarcina* spp. (see S3 Table for a full listing).

All of the finalized models are available in the SBML [363] compatible with COBRApy [230] and can be found at `https://github.com/JosephRyanPeterson/GEMs_methanogens`.

5.3 Results and Discussion

Several comparative genomic (CG) studies of methanogens have been performed to date, yet most have been limited to defining phylogenetic relationships by a restricted examination of methanogenesis genes [12,364,365]; or to comparisons of just a few strains of a methanogenic species [366–369]. One notable exception includes a seminal gut microbiome study where a pan-genome examination of the *Methanobrevibacter smithii* spp. in groups of twins identified a variety of adhesin-life proteins, the function of which was hypothesized to be the creation of diversity in metabolic niches [370]. In another notable study, a targeted examination of closely related *M. mazei* spp. isolates from fresh and marine environments of the Columbia River Estuary revealed differences in primary metabolism to support niche partitioning that affected trimethylamine utilization [214]. The paucity of pan-genomic studies prompted us to examine a broad sampling of species across the *Methanosarcina* genus. To do so, genomes of 27 previously isolated *Methanosarcina* spp. (20 of which were closed) were sequenced and analyzed along with three previously published genomes [14,325,371] (see S1

Appendix Table 5.1 for a full listing of species examined).

We begin our analysis of the results with a broad overview of the pan-genome followed by an examination of the pan-reactome. Subsequently, we examine key metabolic differences. Auxotrophies predicted by the GSMM approach are discussed. Finally, we discuss predictions of new metabolic functions that arise from our combined GSMM/CG approach.

5.3.1 The *Methanosarcina* Pan-Genome is Highly Variable

A pan-genome profile was created using the ITEP toolkit [191] with gene clusters computed via a Monte Carlo clustering approach (using OrthoMCL [352]). Clustering the species by gene family presence/absence yielded five groups consisting of (labelled by typestrain) the *M. barkeri*, *M. calensis*, *M. mazei*, *M. siciliae* and *M. thermophila* (see Figure 5.2A). A *Methanosarcina* core-genome consisting of 1329 genes was identified, along with a pan-genome consisting of 7005 genes (see Figure 5.2B). The *Methanosarcina* pan-genome, which comes in at about 5.3$\times$ its core-genome size is rather modest in light previous studies which identified pan-genomes to core-genome ratios ranging from 1.9 in *Methanobrevibacter smithii* [370], to 5.2 in *Staphylococcus aureus* [372], to 9 in *E. coli* [373], to over 13 in the *Streptococcus* genus [374]. The *Methanosarcina* pan-genome appears to be highly variable, with between 20 and 67% of genes conserved in more than one group (see Fig 5.2b). The pan-genomes of the *M. mazei* and *M. thermophila* groups were largely conserved in other groups. In the case of the *M. thermophila* spp. this is likely due to their small genomes (2400 genes) relative to the other *Methanosarcina* spp. (see Fig 5.2D),

where for the *M. mazei* this might be due to higher propensity for lateral gene transfer arising form their relatively high number of transposases [371]. In the case of the *M. mazei* this may be an adaptation to environmental stresses, as was suggested suggested in a previous study that identified numerous regulatory RNAs antisense to six transposases which were induced by nitrogen stress [148].

Among the core genome, nearly all genes have an associated functional category (Fig 5.2C, top). As expected, the fraction of genes without a putative function increases as the genes become less conserved; however, those with putative function are relatively evenly distributed among functional categories. Of the genes that are found in only one to three organisms, those associated with mobilome (COG category V) constitute the largest fraction (Fig 5.2C, middle). Additionally, genes involved with inorganic ion transport/metabolism (P), defense mechanisms (X), and cell wall/membrane/envelope biogenesis (I & M) constitute large fractions of genes regardless of conservation level. Several large islands of unique genes are completely conserved among each group, ranging between 4 and 11% of the clade-average gene count (Fig 5.2A, B, & D). These islands are of particular interest as they could act "fingerprints" for the unique metabolic functionality associated with each *Methanosarcina* group when examining metagenomic datasets.

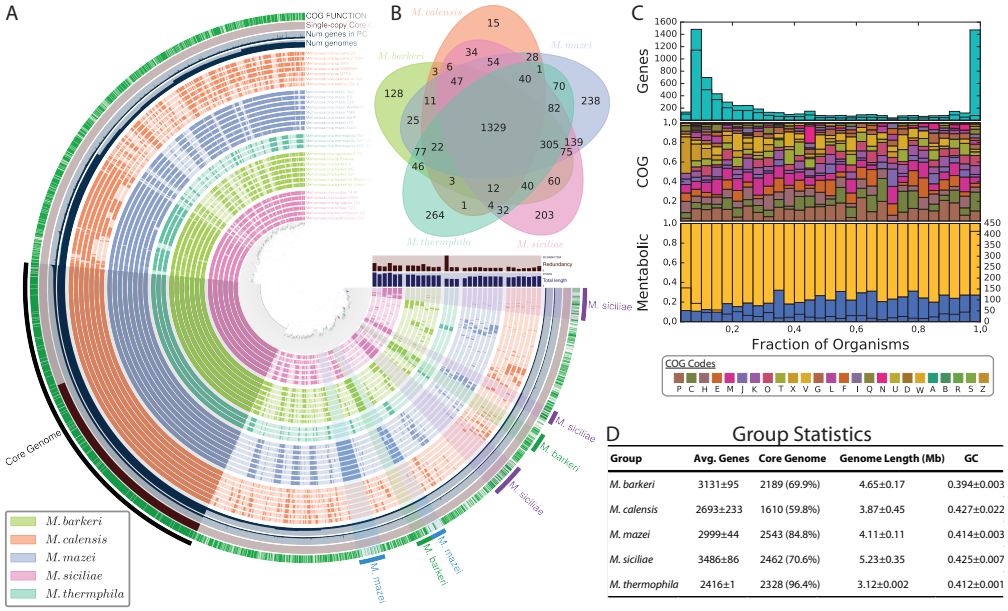
About 20-30% of genes with annotation are associated with metabolic function (*i.e.*, central metabolism, or amino acid, nucleotide, coenzyme, and lipid biosynthesis, *etc.*) regardless of their extent of conservation (see Fig 5.2C,

bottom). All of these observations support the idea that the pan-genome of the *Methanosarcina* genus is highly variable, perhaps as a consequence of the large fraction of mobile elements. Further, among the genes with known function, metabolism is a relatively important fraction. Altogether, this suggests a plethora of metabolic functionality is unaccounted for in current metabolic models.

5.3.2 The *Methanosarcina* Pan-Reactome

We reasoned that genes with a putative function—that is, those assigned with a COG codes—would be candidates for newly discovered function. The high-quality metabolic models available for several *Methanosarcina* spp. [36, 284, 286, 292] could act as a scaffold for newly defined functions; therefore we propagated these models across the 30 species. The benefit of this approach is twofold: 1) the highly-curated manual annotations in these models can be directly applied to each of the organisms, providing more specific information than public databases (*e.g.*, KEGG, UniProt), and 2) incongruences between species could identify diverged metabolic function or highlight gaps of knowledge. This allows the newly discovered pathways/reactions to attach, where possible, to existing pathways and benefit from existing model capabilities.

Models were propagated over each species in an automated fashion using functionality in the ITEP database that was created for the pan-genome analysis (described in Methods Section 5.2.4). Essentially, the gene-reaction-rules for each of the metabolic models were evaluated based on the pres-



Group	Avg. Genes	Core Genome	Genome Length (Mb)	GC
M. barkeri	3131±95	2189 (69.9%)	4.65±0.17	0.394±0.003
M. calensis	2693±233	1610 (59.8%)	3.87±0.45	0.427±0.022
M. mazei	2999±44	2543 (84.8%)	4.11±0.11	0.414±0.003
M. siciliae	3486±86	2462 (70.6%)	5.23±0.35	0.425±0.007
M. thermophila	2416±1	2328 (96.4%)	3.12±0.002	0.412±0.001

Figure 5.2: ***Methanosarcina* Pan-Genome** (A) A visualization of gene clusters demonstrating the diversity of protein families in the *Methanosarcina* spp. examined in this study. Each circular track represents a different species colored by group (*M. siciliae* spp. (pink), *M. barkeri* spp. (light green), *M. thermophila* spp. (dark green), *M. mazei* spp. (blue) and *M. calensis* spp. (orange)). The core genome containing 1329 genes, along with several other clusters of genes that are conserved only in one group, are highlighted by radial wedges. Anvi'o was used to generate the graphic [375]. (B) A Venn diagram showing the overlap in gene content among the different groups. Each ellipse contains the sum total of genes that are completely conserved across the group. (C) (top) Histogram showing the number genes at a given conservation level (cyan) in *Methanosarcina* species, showing the characteristically high level of completely conserved genes and nearly unique genes. The black boxes indicate the fraction of genes with no similarity to a known function (as annotated by COGs). (middle) Fractions of genes by COG category (excluding genes with unknown function). (bottom) Fractions of genes that are involved in metabolic (blue) and non-metabolic (yellow) COG categories. The black histogram indicates count of "metabolic" genes at each fraction of conservation. A legend mapping COG code to color is found below the plots. (D) Statistics derived from the genomes in each of the five species groups.

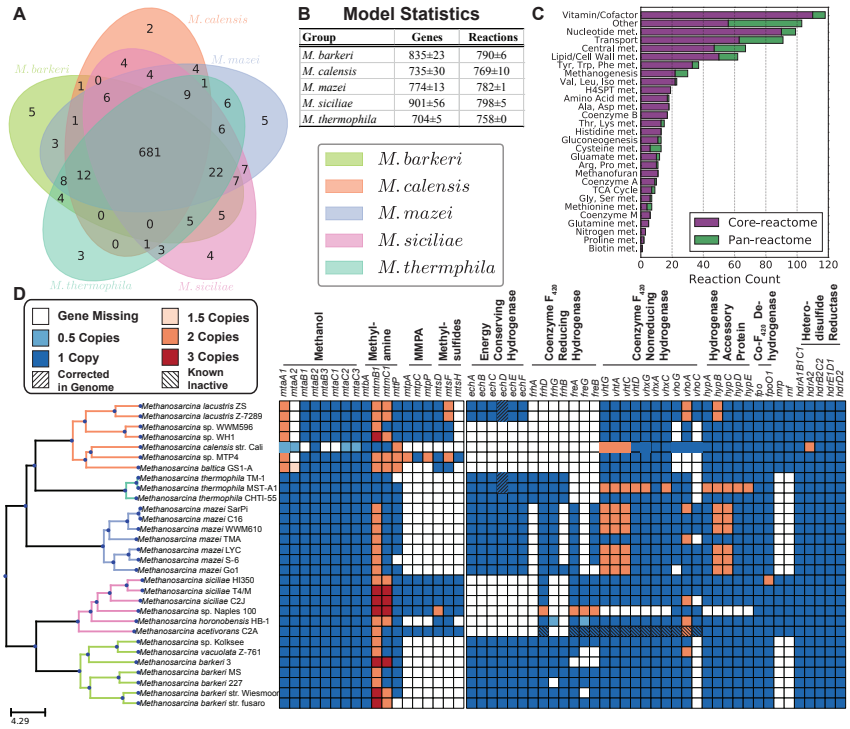


Figure 5.3: **The *Methanosarcina* Pan-Reactome** (A) Venn diagram of model (non-exchange) reactions for each of the five metabolic groups showing a highly conserved core of reactions. (B) Statistics for average gene and reaction counts for the models in each of the five groups. (C) Quantification of core- and pan-reactomes by metabolic subsystem. Reactions are classified as one of 29 metabolic subsystems within the metabolic models. The greatest extent of variability exists for peripheral metabolism such as "Vitamin/Cofactor", "Other", and "Transport", although notable exceptions occur, for example "Cysteine/Methionine" metabolism. (D) A presence/absence table for genes involved in substrate utilization and energy conservation organized by the group clustering. A white square indicates that no genes exist, while dark blue, orange and red squares indicate one, two or three copies of the gene exist in the genome. Light blue and pink squares indicate "0.5" or "1.5" gene copies, meaning that the gene cannot be confidently assigned to a particular homolog (*e.g.*, *mtaC2*/*mtaC3* in *M. calensis* str. Cali). Copies of the *echD* gene were uncalled by automated tools in the *M. calensis* and the *M. thermophila* that were corrected manually. Expression of coenzyme F420 reducing/nonreducing hydrogenases genes are known to be unexpressed in *M. acetivorans* C2A.

ence/absence of homologous genes in the clusters computed by ITEP, to determine if the components necessary to catalyze the reactions exist in the organism of interest. After automatic propagation, most models consisted of between 600 and 700 reactions. These draft models had a number of nonsensical auxotrophies, prompting manual curation (see Results Section 5.3.5). The models generated from this procedure generally do not predict growth, as one or more pathway for an essential biomass component contained gaps. To restore growth, a gap-filling procedure which completes metabolic pathways based on adding the fewest possible reactions was employed. After gap-filling, two rounds of manual curation (described in Methods Section 5.2.5) were performed: 1) identification of nonhomologous genes facilitating one of the gap-filled reactions, and 2) assignment of functions for genes that were completely conserved within each of the five groups. Gene functions were assigned based on COG classification, protein family prediction, or homology to known metabolic genes in other organisms.

Finalized models were significantly larger after manual curation, containing between 750 and 800 reactions associated with between 700 and 900 genes (see Fig 5.3A & B). A core-reactome comprising 681 reactions was identified, constituting between 85% and 90% of the each group's reactome. In contrast to the pan-genome, the pan-reactome was a much smaller percentage of the overall reactome, although each group had at least 2 unique reactions. Notably, a large number of reactions (22) were shared among all organisms except the *M. calensis* group, while another large set of reactions (12) were conserved but for the *M. siciliae* group. The relatively larger

core-reactome compared to core-genome is due to the difficulty in assigning metabolic functionality to genes based on homology in the absence of biochemical or molecular biological evidence. In general, we were unable to assign functions for between 100 and 200 putative metabolic genes in each clade, suggesting a highly variable reactome with rich features still to be determined. Nevertheless, the function for over 210 clusters of genes were assigned in this work, significantly expanding our knowledge of the metabolic capabilities within the *Methanosarcina* genus (see S3 Table for a complete listing of unique additions).

The pan-reactome is heavily biased towards peripheral metabolic pathways (Fig 5.3C). This is readily apparent when conservation is represented on a map of the metabolic reactions (see S1 Appendix Figure 5.8). By far the largest number of non-conserved reactions are in the "Other" metabolic subsystem (as annotated by KEGG/RAST) followed by "Transport". That being sad, nearly 20 central metabolism reactions, and 10 lipid/cell wall reactions exist in the pan-reactome, indicating diversity in sugar and lipid metabolism among the *Methanosarcina* spp. Indeed, the remaining unassigned metabolic genes are replete with functional similarities to glycosyltransferases, UbiA poly-/prenyltransferases, UbiE methyltransferases, UbiG benzoquinol methylases, acyclopropane fatty-acyl-phospholipid synthases, and peptidoglycan deacetylases. The large diversity of membrane lipids [376] and methanochondroitin [377] has long been known in methanogens, and our analysis indicates group-specific differences. The novelty and function of such differences—perhaps they cause species-specific sarci-

nas, or function in phage resistance, such as to Methanosarcina spherical virus [378]—remain to be discovered.

Differences do exist within core metabolic pathways including those involved in biosynthesis of amino acids, nucleotides, vitamins and cofactor (see Fig 5.3C). Of particular interest, are variabilities in cysteine and methionine biosynthesis. Species of the *Methanosarcina* genus generally contain several different pathways to generate cysteine and methionine. As detailed nicely in a recent study by Rauch *et al.* [379], they can generate cysteine either by the ancestral methanogen pathway (*i.e.*, by inserting a sulfide into tRNA-bound phosphoserine to form cysteine), or, more commonly, from *O*-acetyl-L-serine via a sulfhydrylase. Rauch *et al.* discovered that in *M. acetivorans* C2A methionine biosynthesis may occur—analogously to cysteine biosynthesis—via two different pathways: either from aspartate semialdehyde, again via sulfide insertion by a cystatione-β-synthase like protein and its associated ferredoxin (encoded by the genes *MA1821* and *MA1822*, respectively), or, more commonly, from *O*-acetyl-L-homoserine via a sulfhydrylase [379]. In general, the genomes of the *Methanosarcina* spp. contain genes necessary to carry out all four of these pathways; however, several notable exceptions exist (see S1 Appendix Fig 5.12). Strikingly, homologs of *M1821/MA1822* and the genes encoding cysteine synthase (model reaction ID: CYSSr), *O*-acetyl-L-serine acetate-lyase (ACSERHS), homoserine *O*-trans-acetylase (HSERTA), *O*-acetylhomoserine (thiol)-lyase (AHSERL2), and *O*-succinylhomoserine lyase (SHSL2r) are missing from the *M. mazei* spp. (except strain TMA). The lack of these enzymes would require the *M.*

mazei to generate cysteine and methionine using the ancestral pathways (*e.g.*, from phosphoserine and aspartate semialdehyde, respectively). This observation may explain why *M. mazei* encodes three different isoacceptors of $tRNA^{Cys}$ [92, 325]. While it was shown that all three isoacceptors in *M. mazei* Gö1 are functional and that CysRS or SepRS preferentially bind to different isoacceptors [380], our analyses indicate that SepRS is likely the primary source. Further, the *M. mazei* do not contain the SepCysE translation factor, which is essential for efficient formation of cysteine from phosphoserine [381]. The lack of cysteine biosynthesis from L-serine might explain the existence of multiple isoacceptors in these organism; specifically, that multiple isoacceptors help to increase the efficiency of cysteine biosynthesis from phosphoserine in the absense of the translation factor SepCysE. Such examples demonstrate the utility of pan-genomic studies through the lens of GSMM.

We identified group specific differences in methanogenesis genes (Fig 5.3D). Such differences are well conserved among the groups, suggesting that they could be markers for group-specific metabolic functionality in microbial communities. Discussion of nuances in conservation of methanogen genes are deferred to Results Section 5.3.4.

5.3.3 Genome/Reactome Comparison

An interesting analysis that follows from having the pan-genome and pan-reactome is the comparison thereof. To do so, a phylogenetic tree constructed from a concatenated alignment of the core-genome (all 1329 genes) was

compared to results from hierarchically clustering the presence/absence patterns for metabolic reactions (Fig 5.4A). The structure of both trees were highly supported: in the case of the phylogenetic tree, the aLRT branch supports are large (> 4000 for different phylogenetic clades); in the case of the metabolic tree, clades are separated by at least 8 reactions. The results show that an organism's metabolic capabilities need not necessarily coincide with their phylogenetic placement; however, this may be an artifact of propagating metabolic models from *M. acetivorans* C2A and *M. barkeri* str. Fusaro to distant organism such as those in the *M. calensis* group.

In general, the number of reactions in a metabolic model scales with the number of genes in that metabolic model in a highly correlated fashion (Fig 5.4B; Pearson r=0.882, p-value$< 1.2 \times 10^{-10}$). The model gene content follows the genome size (see Fig 5.2D and Fig 5.4B). Similarly, the absolute differences in the number of reactions scales with phylogenetic distance (computed relative to *M. acetivorans* C2A), albeit with a smaller correlation and less significant trend (Fig 5.4C; Pearson r=0.786, p-value$< 7 \times 10^{-7}$). In general, the metabolic groups are very tightly clustered, suggesting that signatures of these groups could be used as markers for metabolic capabilities.

Several significant outliers emerge when comparing phylogenetic and metabolic trees. Of particular note are the three members in the *M. calensis* group that are phylogenetically distant from *M. acetivorans* C2A that break the group into two (Fig 5.4A & C). These organism form a tight metabolic cluster: they group closely with the *M. thermophila* and the *M. mazei*, primarily due to the *absence* pattern of metabolic reactions, rather than the

presence pattern (see Fig 5.5). This split is confounding for the additional reason that in general the *M. calensis* clade are marine organisms, while the *M. mazei* and *M. barkeri* are primarily freshwater. Upon inspection, we discovered that the phylogenetic split coincides with the ability to utilize methyl-mercaptopropionate; the *M. lacustris* strains, *M.* sp. WH1 and *M.* sp. WWM596 possess the necessary genes, while *M. calensis* str. Cali and *M. baltica* GS1-A do not (see Figure 5.3D). This indicates that the metabolic capabilities are not strictly tied to the habitat in which the methanogens are found. A detailed analysis of specific metabolic differences and their implications for the metabolic capabilities of the methanogens, is discussed in the next several section.

5.3.4 Differences in Methanogenesis

It is pertinent at this time to discuss specific differences in methanogenesis between the different metabolic groups, as these differences resemble group structure significantly (Fig 5.3D). With respect to metabolic substrates, the *M. barkeri*, *M. mazei*, and *M. thermophila* are completely devoid of the ability to grow on methyl-mercaptopropionate and methyl-sulfides (although the *M. barkeri* is capable of consuming them via the MtsAB enzymes when grown on acetate [25, 143, 383, 384]). Concurrent with this is the absence of the sodium pumping ferredoxin:methanophenazine oxidoreductase (*rnf*) and multisubunit sodium/proton antiporter (*mrp*) gene clusters, which function in electron transport for acetate grown *Methanosarcina* spp. [104, 296, 385] (Fig 5.3D). Both observations can be attributed to the fact that these species

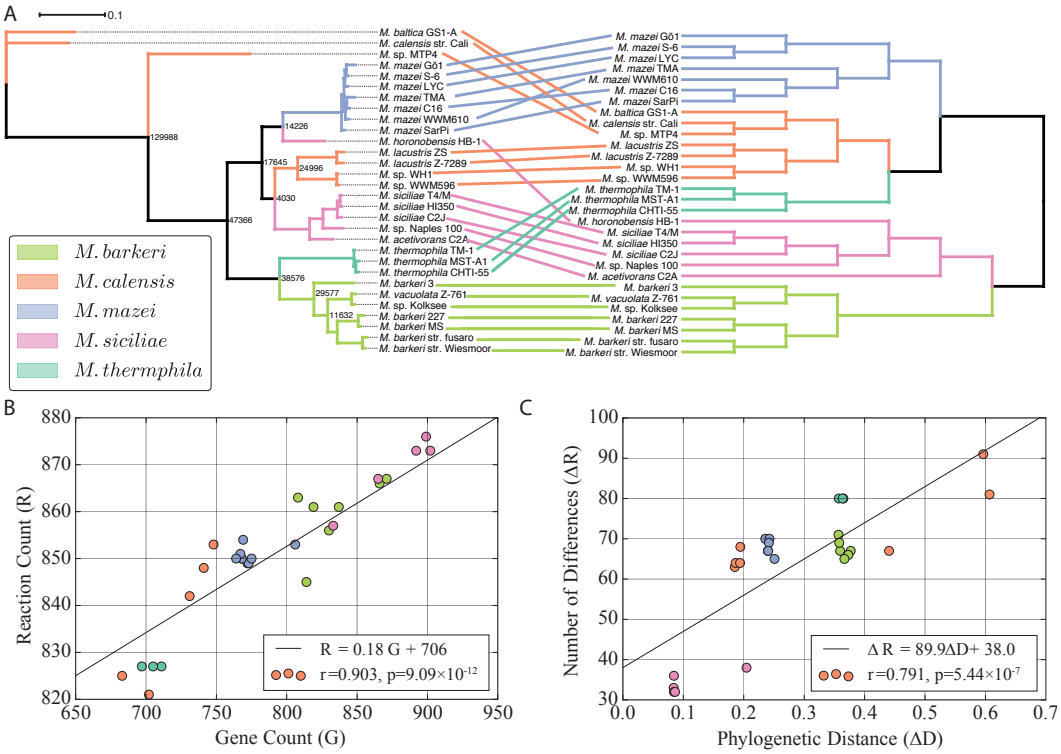


Figure 5.4: **Genome/Reactome Comparison** (A) Trees representing the phylogenetic relationship (left) and metabolic (right) capabilities of the *Methanosarcina* spp. The phylogenetic tree is based on sequence alignment of 1329 conserved genes, while the metabolic tree is based on the conservation pattern of the metabolic reactions with known gene associations. Branch labels indicate the aLRT branch support. The reaction tree was constructed from the metabolic models of the organisms. Five clusters are clearly apparent. The metabolic tree and clusters were generated via hierarchical clustering with Ward's method as the criterion [382]. Interestingly, the position of the *M. thermophila* group is different between the phylogenetic and metabolic trees. (B) Total reaction count, R in the *Methanosarcina* metabolic models is directly correlated to the number conserved metabolic genes, G. (C) The difference gene-associated reactions between the *Methanosarcina* species, ΔR, increases with phylogenetic distance, D. The phylogenetic distance between *M. acetivorans* C2A and the species of interest was calculated from the concatenated alignment of 1329 core conserved genes.

are by-in-large freshwater. It is known that methyl-sulfides and methyl-mercaptopropionate are commonly produced in ocean waters, which would explain the ability of marine organisms to consume them. Further, the Rnf enzyme pumps sodium ions that and are subsequently exchanged for protons by Mrp and methyltetrahydrosarcinapterin:coenzyme M methyl-transferase (Mtr), therefore utilizing the abundance of ions in the ocean. All *Methanosarcina* spp. have genes encoding methanol and methylamine methyltransferases, although the number of paralogs vary (Fig5.3D). Interestingly, several species from different metabolic groups appear to be missing the methylamine transport gene *mttP*, while capable of growth on methylamines. For example, *M. mazei* S-6 and *M. thermophila* TM-1 have been shown to grow on MMA and TMA [326, 329], prompting us to wonder how methylamines are taken up by these species. Another curious feature in the *M. calensis* metabolic group is that rather than having two orthologous methanol coenzyme M:corrinoid methyl-transferase *mtaA* genes, they appear to have two copies of the *mtaA1* paralog (see Fig 5.9A). Evidence indicates that this is due to a gene conversion event. While the sequence of the second copy clearly clusters with the other *mtaA1* homologs (Figure 5.9A), it has the same gene context as the *mtaA2* homolog (see S1 Appendix Fig 5.9B & C). Additionally, the original gene duplication occurred earlier than the branching of the *M. lacustris* spp., as the *mtaA2* paralog is found in all other phylogenetic clades (see S1 Appendix Fig 5.4). In *M. acetivorans* C2A both *mtaA* genes have been shown to support growth on methanol at similar rates when expressed from active promoters [134], although *mtaA2* alone

was insufficient to support growth on methanol in wild type cells [19]. And while the *mtaA2* is expressed in low levels in *M. acetivorans* C2A and *M. thermophila* [19,36,386], it was shown to contribute significantly to methane production [19]. Consequently, what effect—if any—this gene conversion event has on growth is difficult to predict.

Group-specific presence/absence patterns also manifest in the electron transport chain reactions. The *M. calensis* spp. are missing all homologs of the coenzyme F_{420} reducing hydrogenase (*frh*/*fre*) while containing nearly all the homologs for coenzyme F_{420} nonreducing hydrogenase (*vho*/*vht*/*vhx*), requiring these methanogens to use coenzyme F_{420} dehydrogenase (Fpo) and Rnf to regenerate the coenzyme F_{420} pool (see Fig 5.3D). For these reasons, they are not likely to be capable of hydrogenotrophic growth (that is, metabolizing H_2/CO_2) [387]. Strikingly, *M.* sp. Naples 100 appears to be an outlier among the *Methanosarcina* missing genes encoding the hydrogenase accessory protein (*hyp*), an enzyme which introduces the nickel and iron ions during maturation of the iron-nickel hydrogenases [388], along with the *vht*/*vhx* gene clusters. One possible explanation for these discrepancies might be the fact that the genome was not closed, and that the genes exist in the missing regions. Nevertheless, this would be interesting to investigate, as lacking *hyp* would likely render the iron-nickel hydrogenases (*fre* and *vho*) in the organism non-functional.

An interesting conservation pattern among the coenzyme F_{420}-reducing hydrogenaase (*frh*/*fre*; hereafter collectively *frx*) genes, wherein the *M. mazei* and *M. siciliae* orthologs appeared to be a combination of the *frh* and *fre* gene

clusters (see Fig 5.3D), warranting investigation. In general, all *Methanosarcina* spp. contain only a single copy of the *frx* cluster aside from the *M. barkeri* [389]. Within the *fre* cluster in the *M. barkeri*, the maturation protease *D* subunit has been replaced with an *E* gene and the gene cluster is nonessential and insufficient to sustain hydrogenotrophic growth [387]. Conversely, the *frh* cluster in the *M. barkeri* is essential for growth [387]. However, it was unclear to us whether the replacement of the *D* subunit caused the inactivity of the *fre* enzyme. All *M. mazei* spp. are capable of hydrogenetrophic methanogenesis [343, 390], indicating that the apparent *freAB* subunits of coenzyme F_{420}-reducing hydrogenase must be functional. These observation prompted a phylogenetic analysis of the *frx* genes in the *Methanosarcina*. Upon deeper inspection, we discovered that the *freAGB* subunits in the *M. barkeri* are phylogenetically more closely related to the subunits in the *M. siciliae* and the *M. mazei* than are the paralogous *frhAGB* subunits in that species (see S1 Appendix Fig 5.10A). Further, they are more closely related to the orthologs in more distant methanogens. Rather, the *frhAGB* subunits in the *M. barkeri* are closely related to the *frhAGB* subunits in the *M. thermophila* (see S1 Appendix Fig 5.10B). Taken together, this suggests an interesting origin for the two paralogous *frx* gene clusters in the *M. barkeri*. Most likely, the *M. barkeri* obtained a copy of the *M. thermophila frh* via horizontal gene transfer, potentially as a means of repairing the loss of hydrogenotrophic growth due to the deleterious replacements of the *D* with the *E* subunit.

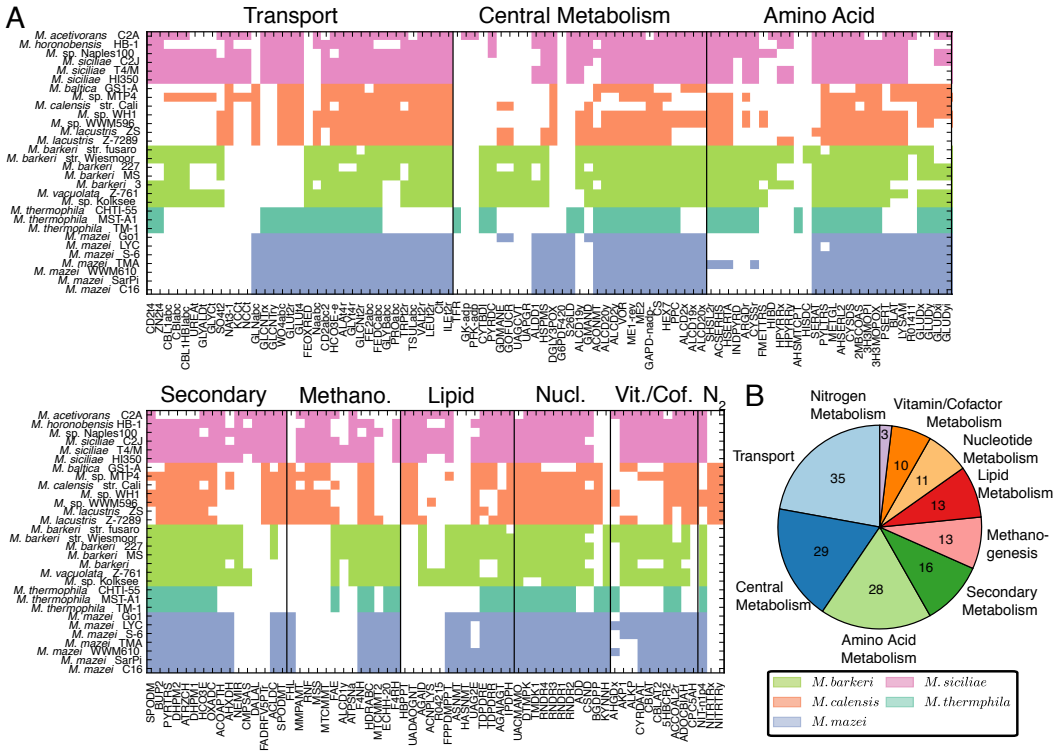


Figure 5.5: **Variability in Reaction Content.** (A) The presence and absence of reactions which are conserved to different extents among the 30 *Methanosarcina* species. Rows represent organisms while columns represent reactions (which are labelled by their model IDs). A mapping of model reaction IDs and the description of the reaction can be found in SI Table 5.4. The ordering of the species mirrors the reaction tree in (Figure 5.4A). Reactions are organized by broad metabolic category. (B) A breakdown of the 158 variable reaction by metabolic subsystem. **Abbreviations: Methano.** - Methanogenesis, **Secondary** - Secondary Metabolism, **Vit./Cof.** - Vitamin and cofactor biosynthesis, **Lipid** - Lipid and cell wall biosynthesis, **Nucl.** - Nucleotide biosynthesis, and $\mathbf{N}_2$ - Nitrogen biosynthesis.

5.3.5 Model Auxotrophies

Pan-reactome analyses coupled with metabolic modeling have the powerful capability of predicting divergence in metabolic capabilities (*i.e.*, auxotrophies, alternative pathway usage, problems in model assumptions). An in-depth analysis of predicted auxotrophies which are not caused by problems in genomes or in the reference metabolic network can reveal the likely existence of entirely missing pathways in the model. To predict auxotrophies, flux balance analysis was run on the models lacking gap-filled, but containing manually curated, reactions. Biomass components were removed one at a time until the model grew. Predicted auxotrophies were examined manually for potential genome annotation issues (*e.g.*, uncalled genes) and model issues (*e.g.*, incorrect assumptions about model biomasses). Results of this process are detailed in Figure 5.6.

Some of these problems could be corrected by performing thorough literature searches or by reviewing model assumptions. As an example of the former, consider the cysteine biosynthesis discussed previously. There are two known pathways for synthesis of cysteine in *Methanosarcina* and many other methanogens: 1) direct synthesis from serine, and 2) tRNA-dependent synthesis from phosphoserine via the SepRS pathway [200] (S1 Appendix Fig. 5.12). *M. barkeri* str. Fusaro and *M. acetivorans* C2A possess both pathways. Though all *Methanosarcina* possess the SepRS pathway, the direct synthesis pathway is missing in *M. mazei* Gö1 [371]. Surprisingly, initial simulations of the models predicted that the bacterial pathway was essential.

Upon further examination, we discovered that the published metabolic networks only included the first step of the characterized SepRS pathway. However, even after fixing this problem, the models still predicted that the Bacterial pathway was essential for growth because the characterized SepRS pathway terminates at cys-tRNA^{Cys} [200] rather than free cysteine. The models assumes that free cysteine is required as a sulfur source for synthesis of several cofactors including Coenzyme A, Coenzyme B, and Coenzyme M. The cysteine requirement for coenzyme A biosynthesis is directly supported in the literature [391] and one of the enzymes in the coenzyme B synthesis pathway is able to utilize cysteine *in vitro* [360]. However, given that many other pathways that depend on cysteine in bacteria have been replaced with non-cysteine dependent counterparts in the methanogens [392,393], and that utilization of cysteine as a sulfur source for coenzymes B and M has not been demonstrated *in vivo* [393], it is conceivable that an alternative unknown mechanism exists for incorporating sulfur into these cofactors. Regardless, comparative modeling has highlighted a gap in knowledge of this pathway that could be a promising target for further characterization.

Analysis of model auxotrophies also revealed problems in lipid biosynthesis. The gene originally annotated as catalyzing the isopentenyl diphosphate hydrolase reaction in the models (reaction ID: IPDPH) was missing in numerous *Methanosarcina* spp. Literature search revealed that a recent characterization of the *M. acetivorans* C2A gene *MA0127* found that it catalyzes the hydroxylation of the a double bond somewhere in hydroxyarchaeol lipid biosynthesis pathway [206]. Further, this gene was found in

all *Methanosarcina* spp., and was thus corrected in the models (Figure 5.6). There is some uncertainty to its attribution to this particular reaction: the authors of the study suggest it functions on a downstream metabolite—perhaps *sn*-2,3-di-*O*-geranylgeranylglycerol phosphate—but do not present direct evidence for the fact [206]. Nevertheless, this examples indicates that the biosynthesis pathway for hydroxyarchaeol lipids may need to be updated in the future as the biochemistry is better characterized.

Genes associated with gap reactions where no homologs to the genes annotated in *M. acetivorans* C2A and *M. barkeri* str. Fusaro could be identified for several reactions. For example, genes were found for the ornithine decarboxylase reaction (reaction ID: ORNDC) in *M. baltica* GS1-A and *M. calensis* str. Cali, and for ATP-dependent Fe^{2+} transporter (FE2abc) in the *M. thermophila* spp. On the other hand, no genes were identified for N-acetylglucosaminylarchaeatidylinositol deacetylase (AGAID) in 19 of the 20 *Methanosarcina* spp. where the reaction was predicted to be essential, highlighting a significant knowledge gap in lipid metabolism. Similarly, genes for the cob(I)yrinic acid a,c-diamide adenosyltransferase (CYRDAAT) reaction necessary for adenosyl-cobalamin biosynthesis, and for coenzyme F_{430} precursor I aminase (F430S1) were missing in 9 and 2 *Methanosarcina* spp. respectively, suggesting the potential for another pathway or perhaps an alternative cofactor.

Other problems required model assumptions to be revised. For example, the recent addition of N-ϵ-acetyl-β-lysine (NABL) to the biomass equation of *M. acetivorans* C2A in our recent paper [36], based on its role in osmotic adap-

Cause	Missing Paths					Biomass		Wrong or Missing Genes				Missing Genes		Unclear						Likely Real								
Auxotrophy	Cysteine	Cysteine	Isoleucine	Methionine	Methionine	Homospermidine	Putrescine	Coenzyme A	Coenzyme M	Coenzyme F_{420}	Other	Lipids	Lipids	ATP	ATP	ATP	ATP	NG: Methyl-Sulfides	NG: MMPA	ATP	Cobalamin	Cobalamin	Deoxythymidine	Coenzyme F_{420}	Lipids	Lipids	Other
Reaction ID	CYSSr	SERAT	METGL	AHSERL2	HSERTA	HSPMS	ORNDC	ASP1DC	PSERT	F430S1	HCO3E	FE2abc	IPDPH	ECHH20	F4RI	RNF	MRP	MTCMMT	MTCMMT	Nat3_1	CBIAT	CYRDAAT	DTMPK	OXADC	AGAIAGT	AGAID	CS
M. baltica GS1-A	!	!	x	x	x	*	!	!	!	x		x	!	x	x				x		x	x			x	x	x
M. calensis Cal i	!	!	x			*	!				!		!	x	x			x	x		x	x			x	x	
M. sp. MTP4	!	!				*		!	!	x			!	x	x										x	x	
M. thermophila CHTI-55			x			*		!	!			!	!			x	x	x	x	x	x	x				x	
M. thermophila MST-A1			x			*		!	!			!	!			x	x	x	x	x	x	x				x	
M. thermophila TM-1			x			*		!	!			!	!			x	x	x	x	x	x	x				x	
M. barkeri 3													!			x	x	x	x		x					x	
M. sp. Kolksee													!			x	x	x	x				x			x	
M. vacuolata Z-761													!			x	x	x	x							x	
M. barkeri 227													!			x	x	x	x								
M. barkeri MS													!			x	x	x	x								
M. bareri Wiesmoor													!			x	x	x	x								
M. barkeri Fusaro																x	x	x	x								
M. lacustris ZS	!							!					!	x	x						x	x		x		!	
M. lacustris Z-7289	!	!						!					!	x	x						x	x				x	
M. sp. WWM596	!	!				*							!	x	x						x	x				x	
M. sp. WH1	!	!				*							!	x	x						x	x				x	
M. acetivorans C2A														?	?												
M. sp. Naples 100											!			x													
M. siciliae HI350														x													
M. siciliae T4/M														x													
M. siciliae C2J														x													
M. hornobensis HB-1														x					x								
M. mazei TMA								!					!			x	x	x	x							x	
M. mazei C16	!				!			!					!			x	x	x	x							x	
M. mazei WWM610	!				!			!					!			x	x	x	x							x	
M. mazei SarPi	!				!			!					!			x	x	x	x							x	
M. mazei Gö1	!				!			!								x	x	x	x							x	
M. mazei S-6	!				!			!								x	x	x	x							x	
M. mazei LYC	!				!			!								x	x	x	x							x	

Figure 5.6: **Auxotrophies** Each column represents a specific auxotrophy caused by missing the genes from the indicated reaction (reaction IDs can be found in S1 Appendix Table 5.4). Different symbols represent different methods of fixing the model in order to resolve inconsistencies between model implications, comparative genomics and physiology. Uncalled genes ("o") were identified after searching for the gene with tBLASTn with an E-value less than 1×10^{-50}. Reactions fixed by adding a new gene or reaction to the model are indicated by a "!", those fixed by revising model assumptions with "*", problems without a clear solution or that represent real auxotrophies by "x". Genes for reactions that are known to be (biochemically) inactive are marked with a "?".

tation [394], resulted in false auxotrophy predictions in many of the other species. This model assumption caused many of the freshwater methanogens, which are missing genes necessary to perform the β-lysine acetyltransferase (BLAT) and lysine 2,3-aminomutase (LYSAM) reactions in NABL biosynthesis, to be unable to grow. This highlights potential pitfalls in pan-reactome approaches when basing observations on highly specific model assumptions, such as those introduced in [36]. The biomass equations of freshwater methanogens were corrected by removing NABL. A full list of auxotrophies and corrections (where possible) are detailed in Figure 5.6.

5.3.6 The Comparative Approach Allows Prediction of New Metabolic Functions

Comparative modeling can also expose candidates for discovery of new pathways that have yet to be fully characterized. During analysis of the completely conserved genes, we discovered seven homologs to those in molybdopterin biosynthesis pathways of other organisms (*i.e.*, *Escherichia coli* and *Pyrococcus furiosus*). Upon inspection of the models we discovered that this pathway had been missing in prior reconstructions. This essential molecules plays a key role in methanogenesis as a cofactor in the formyl-methanofuran dehydrogenase (Fmd) enzyme [395]. During hydrogenotrophic methanogenesis, this enzyme reduces carbon dioxide, attaching it to the methanofuran cofactor as a formyl-moiety using reducing equivalents from reduced coenzyme F_{420}. During acetotrophic/methylotrophic methanogeneis the reverse process is used to generate reducing equivalents that are used to convert

methyl-groups to methane. A pathway, shown in Figure 5.7, was devised based on characterizations of the homologs in these other organisms and was added to the models. The first step of molybdopterin biosynthesis is the cyclization of GTP to (8S)-3′,8-cyclo-7,8-dihydroguanine 5′ triphosphate [396, 397], mediated by the protein MoaA. This is followed by a second cyclization step to form the pyranopterin phosphate "Precursor Z" via the action of MoaC [396, 397]. Subsequently, the enzyme MoaE mediates two thiolation events forming molybdopterin [398]. This step also requires a reaction to regenerate the thiocarboxy moiety of a sulfur carrier protein. Regeneration proceeds in two steps catalyzed by two enzymes. Homologs of these enzymes from *E. coli* [399, 400] were found in all the *Methanosarcina* spp. (*MA0220* and *MA0808* in *M. acetivorans* C2A). Next, the molybdopterin molecule is guanylating by one or both of the MobA and MoaA enzymes [401]. There is some contention in the literature over whether the guanylation step occurs before or after thiolation; however, we have selected the latter. Finally, the MoeA enzyme inserts the molybdate into the molybdopterin to form the complete cofactor [402]. Two homologs of the *moeA* gene exist in all *Methanosarcina* spp., which likely function in generating the molybdate and tungstate versions of the cofactor needed by the molybdenum (*fmd*) and tungsten (*fwd*) versions of the formylmethanofuran dehydrogenases, respectively. We attempted to decipher which homolog was associated with each metal using bioinformatic approaches, but were unsuccessful.

In addition to this molybdopterin biosynthesis, a newly elucidated path-

way for Coenzyme F_{430} was implemented in the models [403, 404] (see S1 Appendix Figure 5.13). The models now contain complete pathways for synthesizing all the methanogenesis cofactors except for methanofuran and methanophenazine. Recent work by the White lab at Virginia Polytechnical Institute elucidated all but the final step of methanofuran biosynthesis [201–204]. While they found that the MfnF enzyme catalyzes the formation of p-(β-aminoethyl) phenoxy-methyl-2-(aminomethyl) furan precursor to methanofuran, this molecule is missing a tail composed of between 6 and 11 glutamate moieties. During the comparative genomic approach, we identified a gene cluster which may perform this function, namely *MA4217-MA4220* in *M. acetivorans* C2A. Several pieces of evidence support this hypothesis. First, all four genes are coexpressed in acetate, methanol and trimethylamine grown cells [34, 36], indicating they likely form an operon. Second, *MA4219* includes a GXGXG domain that is specific to InterPro numbers IPR002489 & IPR017550, the latter of which is found in formylmethanofuran dehydrogenase subunit C flanking the reactive site. Third, the genes flanking *MA4219*, namely *MA4217*, *MA4218* and *MA4220* encode glutamate-pyruvate amidotransferase-like, flavin mononucleotide/FeS cluster containing, and coenzyme F420 hydrogenase subunit beta-like enzymes, respectively. Finally, this gene cluster is conserved among all methanogens and among the *Archaeoglobus* and *Ferroglobus* genuses, both of which are capable of producing methanofuran [405, 406]. For these reasons, we hypothesize that the gene cluster *MA4217-MA4220* (and its homologs) encode the enzyme catalyzing the final step of methanofuran synthesis.

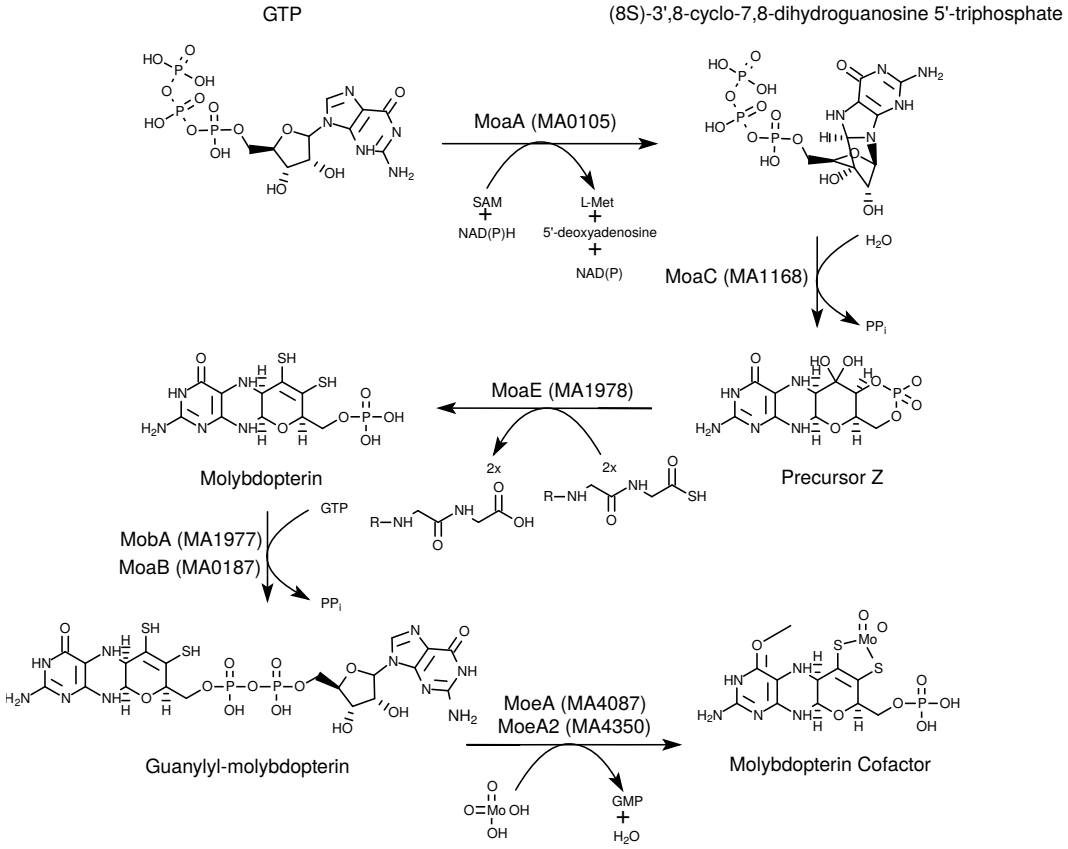


Figure 5.7: **Molybdoterin Coenzyme Biosynthesis Pathway.** The pathway for molybdopterin biosynthesis constructed based on identifying homologous genes to those characterized in other organisms such as *Escherichia coli* and *Pyrococcus furiosus*.

5.4 Conclusion

Herein we utilized a combined comparative genomics (CG)/genome scale metabolic modeling (GSMM) approach to investigate the pan-genome and pan-reactome of 30 species of methanogens of the *Methanosarcina* genus. This work constitutes the first in-depth analysis of the variability of the pan-genome among the *Methanosarcina*, and, more specifically, the vari-

ety of metabolic capabilities embodied in these methanogens. Numerous new predictions about metabolic capabilities and pathways resulted from this study, as well as analyses of specific differences; for example in cysteine/methionine biosynthesis pathways. The resulting information was encoded in 30 new metabolic models that accompany this work. Overall, new characterizations of gene functions significantly expanded our knowledge about methanogen metabolism with 120 newly added genes due to new reactions, 90 new genes in gene-reaction associations, and 52 new metabolites added in aggregate over the five groups of *Methanosarcina* spp. (see S3 Table for a full listing).

GSMM also raised new questions about the growth capabilities, especially with respect to predicted auxotrophies. Do these predictions constitute real auxotrophies or do they point to uncharacterized enzymes? While it is beyond the scope of this study to verify the hypotheses of new genes (*e.g.*, those predicted to be involved in methanofuran and methanophenazine biosynthesis), the fact that targets for further study could be identified demonstrates the utility of our CG/GSMM approach. Two previous studies have utilized similar approaches [311, 372]. These studies showed how such approaches could be used to delineating growth differences caused by adaptations to different microenvironments [311], and to identify strain-specific differences that lead to—and are markers for—pathogenicity of different bacterial species/strains [311, 372]. Our study showed how a combined CG/GSMM approach can be used to predict and characterize new metabolic functions in relatively distantly related species, further demonstrating the

utility of such approaches.

5.5 Supporting Information

5.5.1 Materials and Methods

Computing Conserved Genes and Phylogenies

The ITEP database for *Methanosarcina* spp. described in the main manuscript was used to compute the conservation of genes among the organisms. For each of the gene clusters predicted by ITEP (which correspond to a single gene homolog and its conservation among the organisms in the database), the count of organisms in which the gene is conserved is computed.

A concatenation of the amino acid sequences of 1329 genes that were conserved among all the *Methanosarcina* spp. was aligned using Mafft v7.123b with default parameters. PhyML v20131022 was used to compute a genetic phylogenetic tree from the alignment. The approximate likelihood ratio test (aLRT statistics) was computed and used as a measure of support for the resulting tree.

Model Biomass

Biomass compositions for each of the organisms were estimated from their genome sequence in a fashion similar to that carried out in [284]. The DNA and RNA biomass, along with their ATP requirement and produced pyrophosphate, are found in Table 5.2. Amino acid biomass requirements

can be found in Table 5.3.

Table 5.1: ***Methanosarcina* Strains Studied in this Work** For genomes that are not closed, the number of contigs in the final assembly is cited. The final five are draft genomes that are included for purposes of comparison with the more-complete genomes.

Organism	**Genome**	**Status**	**GenBank ID**
M. acetivorans C2A	[14]	Closed	NC_003552
M. barkeri Fusaro	[325]	Closed	NC_007355
M. mazei Gö1	[371]	Closed	NC_003901
M. baltica GS1-A	This work	33 contigs	
M. barkeri 227	This work	Closed	NZ_CP009530
M. barkeri 3	This work	Closed	NZ_CP009528
M. barkeri MS	This work	Closed	NZ_CP009517
M. barkeri Wiesmoor	This work	Closed	NZ_CP009526
M. calensis Cali	This work	9 contigs	
M. horonobensis HB-1	This work	Closed	NZ_CP009516
M. lacustris Z-7289	This work	Closed	NZ_CP009515
M. lacustris ZS	This work	Closed	
M. mazei C16	This work	Closed	NZ_CP009514
M. mazei LYC	This work	Closed	NZ_CP009513
M. mazei S-6	This work	Closed	NZ_CP009512
M. mazei SarPi	This work	Closed	NZ_CP009511
M. mazei TMA	This work	65 contigs	
M. mazei WWM610	This work	Closed	NZ_CP009509
M. siciliae C2J	This work	Closed	NZ_CP009508
M. siciliae HI350	This work	Closed	CP009507
M. siciliae T4/M	This work	Closed	NZ_CP009506
M. sp. Kolksee	This work	Closed	NZ_CP009524
M. sp. MTP4	This work	Closed	NZ_CP009505
M. sp. Naples 100	This work	5 contigs	
M. sp. WH1	This work	Closed	NZ_CP009504
M. sp. WWM596	This work	Closed	NZ_CP009503
M. thermophila CHTI-55	This work	1 contig	NZ_CP009501
M. thermophila MST-A1	This work	2 contigs	
M. thermophila TM-1	This work	1 contig	NZ_CP009501
M. vacuolata Z-761	This work	Closed	NZ_CP009520

Table 5.2: **Nucleotide Biomass Coefficients.** Coefficients in the biomass equation representing the requisite moles of each nucleotide triphosphate per mole of nucleic acid (e.g. DNA or RNA). Additionally, the amount of ATP consumed, and inorganic phosphate released, during macromolecule synthesis.

	DNA Base (mol/mol DNA)						RNA Base (mol/mol RNA)					
Organisms	**A**	**C**	**G**	**T**	**ATP**	**PPi**	**A**	**C**	**G**	**U**	**ATP**	**PPi**
M. barkeri MS	0.624	2.054	0.403	0.403	2.054	0.624	0.627	1.999	0.388	0.457	1.999	0.527
M. mazei WWM610	0.597	2.054	0.426	0.425	2.054	0.605	0.610	1.999	0.408	0.476	1.999	0.504
M. baltia GS1-A	0.593	2.053	0.458	0.461	2.053	0.541	0.587	1.998	0.448	0.499	1.998	0.464
M. barkeri 3	0.630	2.054	0.402	0.400	2.054	0.621	0.626	2.000	0.387	0.457	2.000	0.530
M. sp. WWM596	0.590	2.054	0.426	0.430	2.054	0.608	0.608	1.999	0.413	0.475	1.999	0.504
M. siciliae HI350	0.584	2.054	0.440	0.445	2.054	0.586	0.593	1.999	0.428	0.486	1.999	0.492
M. mazei C16	0.588	2.054	0.426	0.426	2.054	0.614	0.610	1.999	0.409	0.476	1.999	0.504
M. vacuolata Z-761	0.611	2.054	0.412	0.406	2.054	0.626	0.625	1.999	0.392	0.459	1.999	0.523
M. lacustris ZS	0.598	2.054	0.430	0.427	2.054	0.600	0.610	1.999	0.413	0.474	1.999	0.502
M. mazei TMA	0.601	2.054	0.425	0.423	2.054	0.605	0.612	1.999	0.406	0.475	1.999	0.506
M. horonobensis HB-1	0.599	2.054	0.425	0.423	2.054	0.606	0.611	1.999	0.405	0.472	1.999	0.511
M. thermophila TM-1	0.604	2.054	0.422	0.423	2.054	0.604	0.608	1.999	0.401	0.476	1.999	0.514
M. mazei LYC	0.593	2.054	0.426	0.424	2.054	0.611	0.612	1.999	0.407	0.476	1.999	0.504
M. sp. MTP4	0.557	2.054	0.472	0.471	2.054	0.553	0.564	1.998	0.465	0.515	1.998	0.454
M. mazei G"o1	0.602	2.054	0.426	0.426	2.054	0.600	0.611	1.999	0.407	0.475	1.999	0.505
M. lacustris Z-7289	0.599	2.054	0.430	0.429	2.054	0.596	0.608	1.999	0.415	0.474	1.999	0.502
M. calensis str. Cali	0.538	2.054	0.484	0.487	2.054	0.544	0.555	1.998	0.475	0.523	1.998	0.444
M. sp. Kolksee	0.611	2.054	0.412	0.408	2.054	0.623	0.622	1.999	0.394	0.461	1.999	0.523
M. siciliae TM/4	0.581	2.054	0.439	0.443	2.054	0.590	0.594	1.999	0.427	0.484	1.999	0.493
M. barker 227	0.632	2.054	0.403	0.404	2.054	0.616	0.627	1.999	0.389	0.458	1.999	0.526
M. thermophila CHTI-5	0.604	2.054	0.424	0.423	2.054	0.603	0.606	1.999	0.402	0.477	1.999	0.514
M. mazei S-6	0.600	2.054	0.426	0.425	2.054	0.603	0.612	1.999	0.407	0.475	1.999	0.505
M. mazei SarPi	0.601	2.054	0.426	0.426	2.054	0.601	0.609	1.999	0.409	0.477	1.999	0.504
M. sp. WH1	0.596	2.054	0.429	0.430	2.054	0.599	0.605	1.999	0.415	0.477	1.999	0.501
M. siciliae C2J	0.586	2.054	0.435	0.440	2.054	0.593	0.601	1.999	0.422	0.480	1.999	0.496
M. sp. Naples 100	0.604	2.053	0.421	0.434	2.053	0.594	0.593	1.999	0.428	0.486	1.999	0.493
M. thermophila MST-A1	0.602	2.054	0.425	0.419	2.054	0.609	0.607	1.999	0.402	0.477	1.999	0.514
M. barkeri str. Weismoor	0.621	2.054	0.402	0.405	2.054	0.626	0.627	2.000	0.389	0.456	2.000	0.528

Table 5.3: **Amino Acid Biomass Coefficients.** Coefficients in the biomass equation representing the requisite moles of each amino acid per mole of protein.

	Amino Acid (mol/mol Protein)																			
Organisms	**A**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**	**N**	**P**	**Q**	**R**	**S**	**T**	**V**	**W**	**Y**
M. barkeri MS	0.62	0.11	0.47	0.69	0.38	0.63	0.15	0.69	0.64	0.84	0.21	0.42	0.35	0.23	0.39	0.61	0.48	0.62	0.09	0.32
M. mazei WWM610	0.64	0.11	0.47	0.74	0.38	0.65	0.15	0.68	0.61	0.84	0.22	0.39	0.36	0.22	0.42	0.61	0.46	0.62	0.08	0.32
M. baltia GS1-A	0.63	0.12	0.47	0.77	0.38	0.66	0.16	0.63	0.65	0.86	0.20	0.39	0.36	0.23	0.44	0.56	0.44	0.61	0.08	0.30
M. barkeri 3	0.62	0.11	0.46	0.69	0.38	0.64	0.15	0.70	0.63	0.85	0.21	0.42	0.35	0.23	0.39	0.63	0.49	0.62	0.09	0.32
M. sp. WWM596	0.63	0.11	0.46	0.72	0.39	0.65	0.15	0.68	0.62	0.86	0.22	0.39	0.36	0.23	0.41	0.59	0.46	0.62	0.09	0.31
M. siciliae HI350	0.64	0.11	0.49	0.72	0.39	0.67	0.14	0.66	0.58	0.85	0.21	0.40	0.36	0.22	0.39	0.61	0.49	0.63	0.09	0.33
M. mazei C16	0.63	0.11	0.47	0.73	0.38	0.65	0.15	0.68	0.61	0.84	0.21	0.39	0.36	0.22	0.42	0.61	0.46	0.62	0.08	0.32
M. vacuolata Z-761	0.62	0.11	0.47	0.70	0.38	0.63	0.15	0.69	0.64	0.85	0.21	0.42	0.36	0.23	0.39	0.61	0.48	0.61	0.09	0.32
M. lacustris ZS	0.62	0.12	0.47	0.71	0.39	0.64	0.15	0.67	0.64	0.85	0.22	0.39	0.36	0.23	0.41	0.60	0.47	0.62	0.08	0.31
M. mazei TMA	0.63	0.11	0.47	0.73	0.38	0.65	0.15	0.69	0.61	0.84	0.21	0.39	0.36	0.22	0.42	0.61	0.46	0.61	0.08	0.32
M. horonobensis HB-1	0.63	0.11	0.47	0.72	0.39	0.65	0.15	0.68	0.60	0.86	0.22	0.40	0.35	0.23	0.41	0.62	0.48	0.62	0.09	0.32
M. thermophila TM-1	0.66	0.11	0.46	0.73	0.38	0.65	0.15	0.71	0.61	0.86	0.21	0.38	0.37	0.22	0.43	0.59	0.46	0.62	0.08	0.31
M. mazei LYC	0.63	0.11	0.47	0.74	0.39	0.65	0.15	0.69	0.62	0.85	0.21	0.39	0.36	0.22	0.42	0.60	0.46	0.61	0.08	0.31
M. sp. MTP4	0.67	0.11	0.49	0.78	0.38	0.69	0.15	0.62	0.58	0.87	0.21	0.37	0.37	0.21	0.41	0.59	0.45	0.64	0.09	0.31
M. mazei G"o1	0.63	0.11	0.47	0.73	0.39	0.65	0.15	0.69	0.61	0.85	0.22	0.39	0.36	0.22	0.42	0.61	0.46	0.62	0.08	0.32
M. lacustris Z-7289	0.63	0.11	0.46	0.71	0.39	0.64	0.15	0.68	0.64	0.86	0.22	0.39	0.35	0.23	0.41	0.59	0.46	0.62	0.08	0.31
M. calensis str. Cali	0.69	0.11	0.47	0.78	0.38	0.71	0.15	0.59	0.61	0.89	0.20	0.36	0.38	0.20	0.44	0.59	0.44	0.63	0.09	0.30
M. sp. Kolksee	0.62	0.11	0.47	0.70	0.38	0.63	0.15	0.69	0.63	0.85	0.21	0.42	0.36	0.23	0.39	0.61	0.48	0.61	0.09	0.32
M. siciliae TM/4	0.63	0.11	0.49	0.72	0.39	0.66	0.14	0.66	0.58	0.85	0.21	0.40	0.36	0.22	0.39	0.61	0.49	0.63	0.09	0.33
M. barker 227	0.62	0.11	0.47	0.70	0.38	0.63	0.15	0.69	0.64	0.85	0.21	0.42	0.35	0.23	0.38	0.62	0.48	0.61	0.09	0.32
M. thermophila CHTI-5	0.66	0.11	0.46	0.73	0.38	0.65	0.15	0.71	0.60	0.86	0.21	0.38	0.37	0.22	0.43	0.59	0.46	0.62	0.08	0.31
M. mazei S-6	0.63	0.11	0.47	0.74	0.38	0.65	0.15	0.69	0.61	0.85	0.21	0.39	0.36	0.22	0.42	0.61	0.46	0.61	0.08	0.31
M. mazei SarPi	0.64	0.11	0.47	0.74	0.38	0.65	0.15	0.69	0.61	0.85	0.22	0.39	0.36	0.22	0.42	0.61	0.46	0.62	0.08	0.31
M. sp. WH1	0.63	0.11	0.46	0.73	0.39	0.65	0.15	0.68	0.62	0.86	0.22	0.38	0.36	0.23	0.41	0.59	0.46	0.62	0.08	0.31
M. siciliae C2J	0.62	0.11	0.49	0.72	0.39	0.65	0.15	0.66	0.60	0.84	0.21	0.41	0.36	0.22	0.40	0.61	0.49	0.62	0.09	0.33
M. sp. Naples 100	0.63	0.11	0.49	0.72	0.39	0.67	0.14	0.66	0.57	0.84	0.21	0.40	0.36	0.22	0.39	0.62	0.50	0.63	0.09	0.33
M. thermophila MST-A1	0.66	0.11	0.46	0.73	0.37	0.65	0.15	0.71	0.60	0.86	0.21	0.38	0.37	0.22	0.43	0.59	0.46	0.62	0.08	0.31
M. barkeri str. Weismoor	0.61	0.11	0.47	0.69	0.38	0.63	0.15	0.69	0.63	0.84	0.21	0.43	0.35	0.23	0.38	0.62	0.49	0.61	0.09	0.33

5.5.2 Results and Discussion

Conserved Genes

The OrthoMCL approach identified 7005 unique gene clusters. Of these 1329 were conserved in all of the *Methanosarcina*, while only 134 were unique to only a single specie. Up to 4458 genes were conserved in less than half of the species, indicating that only 2547 were conserved in at least half of the organisms.

Table 5.4: **Variably Conserved Reactions.** A listing of reactions which are conserved to differing extents among the *Methanosarcina* spp. and shown in Figure 5.5. The reaction ID is from the metabolic model along. The function of each reaction is briefly presented in the description.

Reaction ID	Reaction Description
CD2t4	Cadmium transport out via antiport
ZN2t4	zinc transport out via antiport
CBL1abc	Cob(1)alamin transport via ABC system
CBIabc	Cobinamide transport via ABC system
CBL1HBIabc	Cob(1)alamin-HBI transport via ABC system
UREAt	urea transport via facilitate diffusion
GLYALDt	glyceraldehyde facilitated diffusion
GLYCt	glycerol transport via channel
SO4t2	sulfate transport in via proton symport
NAt3-1	sodium proton antiporter
KCCt	K+-Cl- cotransport
NCCt	Na+-Cl- cotransport
GLNabc	Glutamine ABC transporter
GLCNTrx	D-Gluconate:NAD+ 5-oxidoreductase

Table 5.4 (cont.)

Reaction ID	**Reaction Description**
GLCNTry	D-Gluconate:NADP+ 5-oxidoreductase
WO4abc	Tungstenate export by ABC transport
GLUt2r	L-glutamate transport via proton symport, reversible
GLUt4	Na+/glutamate symport
FEOXRED	Fe(II):oxygen oxidoreductase ([FeO(OH)]core-producing)
Naabc	Na-exporting ATPase
CD2abc2	cadmium transport in via ABC system
HCO3E-e	HCO_3^{2-} equilibration reaction
ALAt4r	Alanine-Sodium symporter
GLYt4r	glycine reversible transport via sodium symport
GLCNt2r	D-gluconate transport via proton symport, reversible
FE2abc	iron (II) transport via ABC system
FEDCabc	iron (III) dicitrate transport via ABC system
GLYBabc	Glycine betaine transport via ABC system
PROabc	L-proline transport via ABC system
TRPt2r	L-tryptophan reversible transport via proton symport
TSULabc	thiosulfate transport via ABC system
VALt2r	L-valine reversible transport via proton symport
LEUt2r	L-leucine reversible transport via proton symport
Clt	chlorideion transport out via diffusion
ILEt2r	L-isoleucine reversible transport via proton symport
TFR	Thiol:fumarate reductase
GK-adp	ADP specific glucokinase
PFK-adp	ADP specific phosphofructokinase
CYTBDI	ubiquinol:O2 oxidoreductase (electrogenic, non H+-transporting)
PYRDC	pyruvate decarboxylase
GDMANE	GDP-4-dehydro-6-deoxy-D-mannose epimerase

Table 5.4 (cont.)

Reaction ID	Reaction Description
GOFUCR	GDP-4-oxo L-fucose reductase
UAGCVT	UDP-N-acetylglucosamine-1-carboxyvinyltransferase
UAPGR	UDP-N-acetylenolpyruvoylglucosamine reductase
ALDD1	aldehyde dehydrogenase (formaldehyde, NAD)
HSPMS	homospermidine synthase
DGLY3POX	D-glyceraldehyde-3-phosphate:ferredoxin oxidoreductase
G6PDF420	coenzyme F420-dependent glucose-6-phosphate dehydrogenase
S26LD	L-sorbosone dehydrogenase
ALCD19y	alcohol dehydrogenase (glycerol, NADP)
GMAND	GDP-D-mannose dehydratase
ACONMT	Trans-aconitate methyltransferase
ALCD20y	alcohol dehydrogenase (2-propanol) (NADP)
ALCD2y	alcohol dehydrogenase (ethanol, NADP)
VOR	2-oxoisovalerate ferredoxin reductase
ME1-rev	malic enzyme
ME2	malic enzyme (NADP)
GAPD-nadp-	glyceraldehyde 3-phosphate dehydrogenase (NADP)-phosphorylating
CS	citrate synthase
HEX7	hexokinase (D-fructose:ATP)
PPC	phosphoenolpyruvate carboxylase
ALCD2x	alcohol dehydrogenase (ethanol)
ALCD19x	alcohol dehydrogenase (glycerol, NAD)
ALCD20x	alcohol dehydrogenase (2-propanol) (NAD)
SHSL2r	O-succinylhomoserine lyase (H2S)
ACSERHS	O3-Acetyl-L-serine acetate-lyase (adding hydrogen sulfide)
HSERTA	homoserine O-trans-acetylase
INDPYRD	Indole-3-pyruvate decarboxylase

Table 5.4 (cont.)

Reaction ID	**Reaction Description**
ARGDr	arginine deiminase
CYSSr	cysteine synthase
FMETTRS	methionyl-tRNA formyltransferase
HIBD	3-Hydroxy-2-methylpropanoate:NAD+ oxidoreductase
HPYRRx	Hydroxypyruvate reductase (NADH)
HPYRRy	Hydroxypyruvate reductase (NADPH)
AHSMTCPT	O-acetyl-L-homoserine:methanethiol 3-amino-3-carboxypropyltransferase
HISDC	histidine decarboxylase
SERAT	serine O-acetyltransferase
PYLTRS	pyrrolysyl-tRNA synthase
METGL	methionine g-lyase
AHSERL2	O-acetylhomoserine (thiol)-lyase
CYSDS	Cysteine Desulfhydrase
2MBCOAS	2-methylbutyrate CoA synthesis
3H3MOPI	(S)-2-aceto-2-hydroxybutanoate:$NADP^+$ oxidoreductase (isomerizing)
3H3MOPOX	(R)-2,3-dihydroxy-3-methylpentanoate:$NADP^+$ oxidoreductase
PSERT	phosphoserine transaminase
BLAT	beta-lysine acetyltransferase
LYSAM	lysine 2,3-aminomutase
R01411	5-Methylcytosine aminohydrolase
GLUDx	glutamate dehydrogenase (NADH)
GLUDy	glutamate dehydrogenase (NADPH)
GLUDxi	glutamate dehydrogenase (NAD)
GLUDyi	glutamate dehydrogenase (NADP)
SPODM	superoxide dismutase
BUP2	β-ureidopropionase (D-3-amino-isobutanoate forming)
PYRTRS	pyrrolysyl-tRNA synthase

Table 5.4 (cont.)

Reaction ID	**Reaction Description**
DHPM2	dihydropyrimidinase dihydrothymine
ATRZCH	atrazine chlorohydrolase
DHPM1	5,6-dihydrouracil dihydropyrimidinase
HCO3E	HCO3 equilibration reaction
OXADC	oxalate decarboxylase
ACOAPTH	phosphinothricin acetyltransferase
AHXDH	6-aminohexanoate dimer hydrolase
NEMIR	N-ethylmaleimide reductase
CMPSAS	CMP sialic acid synthase
DALAL	D-alanine-D-alanine ligase
FADRFV5PTr	O-acetyl-L-homoserine:methanethiol 3-amino-3-carboxypropyltransferase
ACLDC	acetolactate decarboxylase
SPODMT	superoxide dismutase
FHL	Formate hydrogenlyase
MMPAMT	Methylmercaptopropionate:coenzyme M methyltransferase
RNF	methanophenazine reductase
MSS	methylsulfide synthase (Mts?)
MTCMMT	Methylthiol: coenzyme M methyltransferase
FAE	formaldehyde-activating enzyme
ALCD1y	alcohol dehydrogenase (methanol)
ATPSNa	Sodium-transporting ATPase
F4NH	Coenzyme F_{420} nonreducing Hydrogenase
HDRABC	Heterodisulphide reductase ABC
MTCMMT2	Methylthiol:coenzyme M methyltransferase
ECHH-20	energy-conserving hydgrogenase (Ech) Hydrogenase
F4RH	Coenzyme F_{420} reducing hydrogenase
HBPPT	Bis(5′-nucleosyl)-tetraphosphatase (asymmetrical)

Table 5.4 (cont.)

Reaction ID	Reaction Description
UADAOGNT	UDP-4-amino-4-deoxy-L-arabinose:oxoglutarate aminotransferase
AGAID	N-acetylglucosaminylarchaeatidylinositol deacetylase
ACNPLYS	N-acetylneuraminate synthase;
R04215	CTP:N-glycoloylneuraminate cytidylyltransferase
FPPDMPPT	farnesyl diphosphate:dimethylallyldiphosphate prenyltransferase
ASNMT	Phosphatidylethanolamine N-methyltransferase
HASNMT	Phosphatidylethanolamine N-methyltransferase
UAG2E	UDP-N-acetylglucosamine 2 epimerase
TDPDRE	dTDP-4-dehydrorhamnose 3,5-epimerase
TDPDRR	dTDP-4-dehydrorhamnose reductase
AGAIAGT	archaeatidylinositol N-acetylglucosaminyltransferase
IPDPH	IPDP hydrolase
UACMAMO	UDP-N-acetyl-D-mannosamine oxidoreductase
DTMPK	dTMP kinase
TMDK1	thymidine kinase (ATP:thymidine)
RNDR4	ribonucleoside-diphosphate reductase (UDP)
RNDR3	ribonucleoside-diphosphate reductase (CDP)
RNDR1	ribonucleoside-diphosphate reductase (ADP)
RNDR2	ribonucleoside-diphosphate reductase (GDP)
ADD	adenine deaminase
CSND	Cytosine deaminase
BGDPP	Bis(5′-nucleosyl)-tetraphosphatase (asymmetrical)
KYNNH	N-Formyl-L-kynurenine amidohydrolase
AHGDx	(S)-α-hydroxyglutarate dehydrogenase
AKP1	alkaline phosphatase (Dihydroneopterin)
ALKP	alkaline phosphatase
CYRDAAT	cob(I)yrinic acid a,c-diamide adenosyltransferase

Table 5.4 (cont.)

Reaction ID	**Reaction Description**
CBIAT	Cobinamide adenyltransferase
CBLAT2	cob(I)alamin-HBI adenosyltransferase
5HBCR2	oxidized 5-hydroxybenzimidazolylcob(I)amide reduction
ACCOAL2r	acetate-CoA ligase (ADP-forming)
ADOCBIAH	Adenosyl cobinamide amidohydrolase
CPC5AH	precorrin-5A reductase
NIT-n1p4	nitrogenase
NITRTRx	Nitrite reductase
NITRTRy	Nitrite reductase

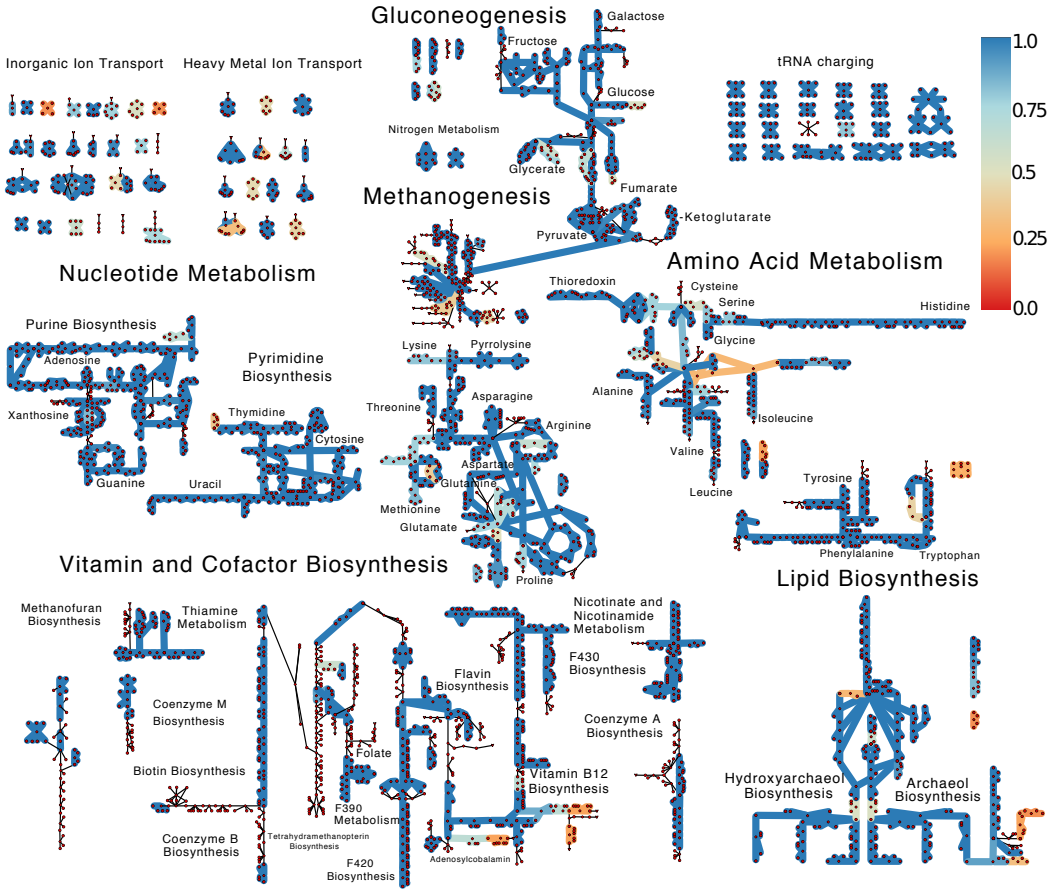


Figure 5.8: **Mapping of Reaction Conservation.** The extent of reaction conservation represented on a map of the reaction network in the *M. acetivorans* C2A model *i*ST807 [36]. Reactions and metabolites are indicated by green diamonds and red circles, respectively, with dependencies between of the reaction on the metabolites indicated by edges connecting nodes. The fraction of organisms that contain the reaction are indicated by the color of these edges with more highly (lowly) conserved reactions indicated by cooler (warmer) colors.

Table 5.5: **Conservation Statistics in *Methanosarcina* Clades** .

Clade	Fully Conserved	Unique	Average Gene Count
M. calensis	1610 (59.8%)	3	2693±233
M. barkeri	2189 (69.9%)	35	3131±95
M. mazei	2543 (84.8%)	36	2999±44
M. siciliae	2462 (70.6%)	3	3486±86
M. thermophila	2328 (96.4%)	98	2416±1

Phylogenies

A concatenated gene alignment of the 1329 completely conserved genes was used to construct a phylogenetic tree (see Fig. 5.4a). The resulting tree had an overall support (aLRT statistic [407]) of 4568821. A comparison of the phylogenetic tree to the tree constructed from conserved metabolic capability (see Fig. 5.4a) reveals two interesting features. First, that the position of the *M. thermophila* clade clusters differently between the two trees, indicating that it is genetically closer to *M. barkeri* while more metabolically similar to *M. calensis* and *M. mazei*. Second, that the *M. calensis* clade is actually a combination of phylogenetically distant organisms.

Hypothesized Methanophenazine Biosynthesis Pathway

As a final example of how the comparative approaches can generate hypotheses about missing new metabolic functions, we attempted to identify genes that could potentially play a role in methanophenazine biosynthesis. Five genes conserved in all the *Methanosarcina* spp. were identified, primarily based on the function and ability to bind a long isoprenoid chain. Following the general patterns in known phenazine biosynthesis [408], the pathway in

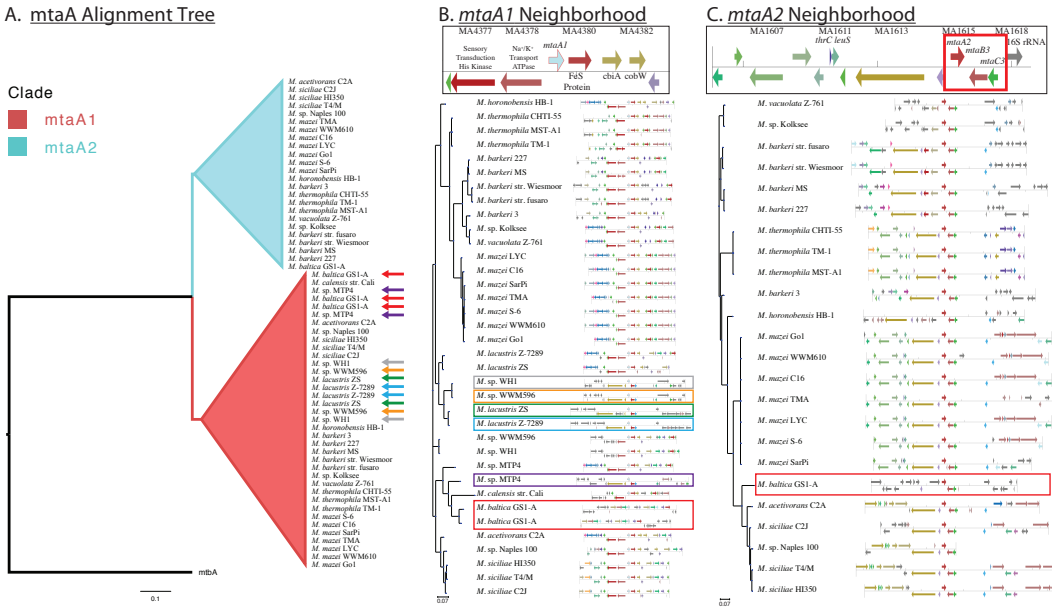


Figure 5.9: **Gene Tree for Methanol *mtaA1* Homologs** A) Gene tree constructed from sequences of the methanol coenzyme M:corrinoid methyltransferase homologs *mtaA1* and *mtaA2* from the*Methanosarcina* spp. showing that all the *mtaA* genes in the *M. calensis* metabolic clade cluster within the *mtaA1* clade (indicated by colored arrows). The dimethylamine coenzyme M:corrinoid methyltransferase mtbA gene is used to root the tree. Phylogenetic trees of the *mtaA1* and *mtaA2* genes identified in the tree of (A) along with their gene context are shown in (B) and (C). At the top of these figures are representative gene contexts for each *mtaA* homolog. Genes indicated by arrows in (A) are singled out with similarly colored boxes. Many of the apparent *mtaA1* homlogs for species the *M. calensis* metabolic clade clearly have the gene context of the *mtaA2* gene. These support the conclusion of gene conversion discussed in the main text.

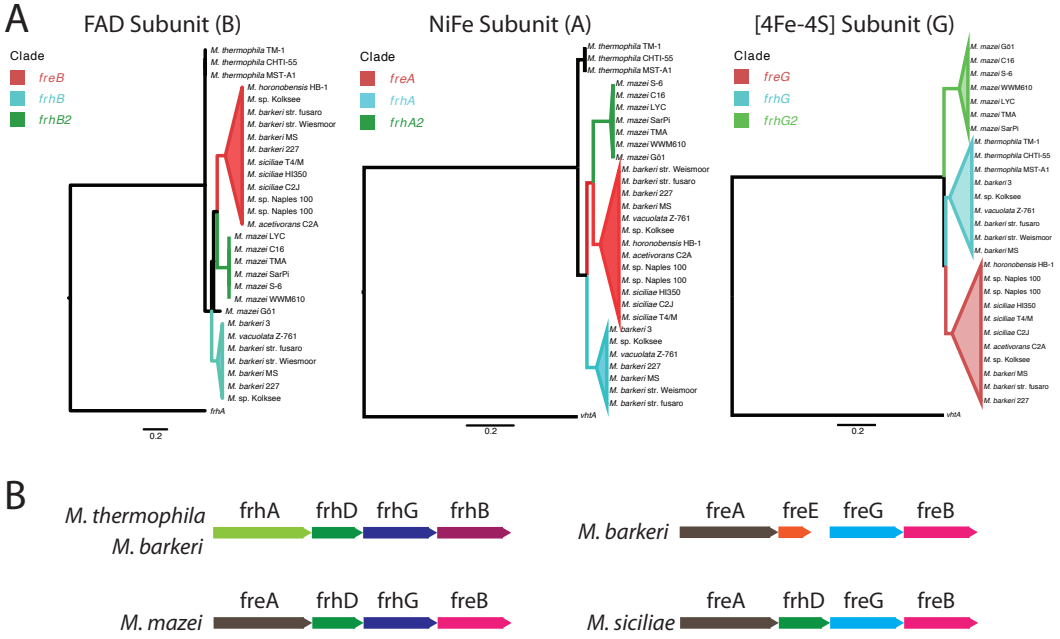


Figure 5.10: **Phylogeny of Coenzyme F420 Reducing Hydrogenase Subunits**. A) Evolutionary relationships between the homologs of the FAD binding (left), catalytic (middle) and 4Fe-4S (right) subunits of the coenzyme F420 reducing hydrogenase (known as subunits B, A and G, respectively). The trees demonstrate that the three subunits of the enzyme in the *M. siciliae* genes are more similar to the *fre* subunits from *M. barkeri*. B) Four gene clusters based on the trees in (A) showing the inferred relationship between the four subunits using the naming scheme from [389]. Genes are colored by similarity, but independently from the colors in (A).

Fig 5.11 is hypothesized. It proceeds in five steps, starting with the phosphorylation of fasenyl geranol via a dolichol kinase, followed with attachment of phenol via benzoate polyprenyltransferase. An amine, perhaps donated from a glutamine, necessary for the formation of the heterocyclic ring structure of the phenazine. The next step might be catalyzed by a dihydropteroate synthase analog using 6-amino-5-oxocyclohex-2-ene-1-carboxylic acid. This structure would then need reduced to form the dihydromethanophenazine, perhaps by a NADH peroxidase/2,4-dienoyl-CoA reductase homolog. These last two examples illustrate how a combined pan-genome/pan-reactome approach can be used to generate hypotheses about new metabolic functions.

Model Improvements

Genes annotated with COG categories related to metabolism (*i.e.*, amino acid, nucleotide, carbohydrate, coenzyme biosynthesis *etc.*) that were fully conserved among all the *Methanosarcina* spp. were enumerated. Based on putative arCOG annotations and homology comparisons the metabolic roles of many of these genes were proposed. When possible, genes were linked with existing model reactions which were missing gene associations, before suggesting additional reactions to be added to the model. From about 150 fully conserved genes, about 30 could be assigned a putative function, demonstrating how a combined genomic/modeling approach can be mutually correcting. In the following paragraphs the proposed functions are described/diagrammed. Additional information about how new functions were annotated as well as a list of all additions can be found in S1 Table 2.

Molybdopterin Cofactor Synthesis (New) Molybdopterin is an important cofactor in the methanogenesis enzyme Fmd; however, the molybdopterin biosynthetic pathways were missing from prior methanogen reconstruction. Six enzymes that are homologous to those in the pathway in *E. coli* [396–398, 401, 402] were identified as being conserved in all the *Methanosarcina* spp. The genes encode proteins that perform all but one of the biosynthetic steps; thus, their addition to the model represents a significant advance in completeness. The reactions include:

1) Cyclyzation of GTP with reducing equivalents derived from NAD(P)H in accordance with that in *E. coli* [396, 397].

Reaction:

$$GTP + S\text{-}adenosyl\ methionine + NAD(P)H \xrightarrow{\text{MA0105}}$$
$$(8S)\text{-}3', 8\text{-}cyclo\text{-}7, 8\text{-}dihydroguanine\ 5\prime\ triphosphate + NAD(P) + L\text{-}methionine$$

Model Reaction (MOCO1):

$$gtp[c] + amet[c] + nadh[c] \xrightarrow{\text{MA0105}} 38c78dhg5tp[c] + nad[c] + met - L[c] + dad - 2[c]$$

2) Cyclization to form pyranopterin monophosphate as in *E. coli* [396, 397].

Reaction:

$$(8S)\text{-}3', 8\text{-}cyclo\text{-}7, 8\text{-}dihydroguanine\ 5'\ triphosphate + H_2O \xrightarrow{\text{MA1168}}$$
$$Precursor Z + diphosphate$$

Model Reaction (MOCO2):

$$38c78dhg5tp[c] + h2o[c] \xrightarrow{\text{MA1168}} precZ[c] + ppi[c]$$

3) Thiolation by a pyranopterin phosphate sulfurtransferase as in *E. coli* [398].

Reaction:

$$Precursor Z + 2\ Thiocarboxy\text{-}[sulfur\text{-}carrierprotein] \xrightarrow{\text{MA1978}} Molybdopterin + 2\ Sulfur\text{-}carrierprotein$$

Model Reaction (MOCO3):

$$preZ[c] + (2)thiocarboxy[c] \xrightarrow{\text{MA1978}} mopterin[c] + (2)carboxy[c]$$

Necessitating also:

$$cys\text{-}L[c] + carboxy[c] \xrightarrow{\text{MA0220/MA0808}} ala\text{-}L[c] + thiocarboxy[c]$$

This reaction has been shown to to proceed via a two step process, each mediated by a separate enzyme, in *E. coli* [399,400]. Interestingly, homologs of both enzymes were conserved among the *Methanosarcina* spp. supporting the idea that a similar mechanism is employed. In the first step the cysteine desulfurase enzyme IscS (encoded by MA0808) transfers a sulfur to the formate dehydrogenase accessory protein FdhD (encoded by MA0220) [399]. In the second step, a dimer of FdhD activates the molybdopterin cofactor by transferring two sulfurs to the cofactor (which ultimately bind the molybdenum metal), as evidenced by co-crystals containing the FdhD enzyme in complex with molybdopterin cofactor [400].

4) Guanylylation by an a MoaB or MobA analog which has been shown to

act in adenylylation of molybdopterin in *Pyrococcus furiosus* [401].

Reaction:

$$Molybdopterin + GTP \xrightarrow{\text{MA0187/MA1977}} Guanylyl\text{-}molybdopterin + diphosphate$$

Model Reaction (MOCO4):

$$mopterin[c] + atp[c] \xrightarrow{\text{MA0187/MA1977}} gmopterin[c] + ppi[c]$$

5) And finally, insertions of the molybdenum ion as in *E. coli* [402].

Reaction:

$$Guanylyl\text{-}molybdopterin + Molybdate \xrightarrow{\text{MA4350/MA4087}} Molybdopterin\ coenzyme + GMP + 2H_2O$$

Model Reaction (MOCO5):

$$gmopterin[c] + mobd[c] \xrightarrow{\text{MA4350/MA4087}} mopterinco[c] + gmp[c] + (2)h2o[c]$$

Coenzyme F_{430} Biosynthesis (New) Recently, two research groups independently discovered the gene cluster that converts sirohydrochlorin into coenzyme F_{430} [403, 404]. These studies identified five steps in the pathway, four of which are catalyzed by five proteins encoded by a cluster named cfbAECDB (the nomenclature of the earlier publication is adopted here [403]). The reactions can be seen in Figure 5.13 and are encoded in the model as:
1) Insertion of a nickel ion into the porphyrin ring.

Reaction:

$$Sirohydrocholin + Ni^{2+} \xrightarrow{\text{MA3631}} Ni^{2+}\text{-}sirohydrochlorin + 2H^{+}$$

Model Reaction (CFB1):

$$scl[c] + ni2[c] \xrightarrow{MA3631} niscl[c] + 2h[c]$$

2) Amidation.

Reaction:

$$Ni^{2+}\text{-} - sirohydrochlorin + 2\ ATP + 2\ \text{L-}Gln + 2H_2O \xrightarrow{\text{MA3626}}$$

$$Ni^{2+}\text{-}sirohydrochlorin\ a,c\text{-}diamide + 2\ ADP + 2\ \text{L-}Glu + 2orthophosphate$$

Model Reaction (CFB2):

$$niscl[c] + (2)atp[c] + (2)gln - L[c] + (2)h2o[c] \xrightarrow{\text{MA3626}} niscldnh2[c] +$$

$$(2)adp[c] + (2)glu - L[c] + (2)pi[c]$$

3) Hydration.

Reaction:

$$Ni^{2+}\text{-}sirohydrochlorin\ a,c - diamide + 6\ ATP + 6\ H_2O + 7\ H^{+}$$

$$\xrightarrow{\text{MA3627/MA3628}}$$

$$Ni^{2+}\text{-}hexahydrosirohydrochlorin\ a,c\text{-}diamide + 6\ ADP + 6\ orthophosphate$$

Model Reaction (CFB3):

$$niscldnh2[c] + (6)atp[c] + (6)h2o[c] + (7)h[c] \xrightarrow{\text{MA3627/MA3628}}$$

$$ni6hscldnh2[c] + (6)adp[c] + (6)pi[c]$$

4) Spontaneous reaction.

Reaction:

$$Ni^{2+}\text{-}hexahydrosirohydrochlorin\ a,c\text{-}diamide \rightarrow seco\text{-}F_{430}$$

Model Reaction (CFB4):

$$ni6hscldnh2[c] \rightarrow secof430[c]$$

5) Coenzyme f430 synthase.

Reaction:

$$seco\text{-}F_{430} + ATP \xrightarrow{\text{MA3630}} Coenzyme\ F_{430} + ADP + Pi + H_2O$$

Model Reaction (CFB5):

$$secof430[c] + atp[c] \xrightarrow{\text{MA3630}} f430[c] + adp[c] + pi[c] + h2o[c]$$

Soluble Heterodisulfide Reductase Homolog 2 (New) A recent study filled a large knowledge gap by characterizing the widely conserved second soluble heterodisulfide homolog (HdrA2B2C2) from *M. acetivorans* C2A [409]. They found dual functionality for the enzyme complex:

1) Heterodisulfide reduction with reduction via ferredoxin:

Reaction:

$$2\ H_2\text{-}Ferredoxin + CoM - CoB \xrightarrow{\text{MA2868/MA4236/MA4237}} CoM + CoB + 2\ Ferredoxin$$

Model Reaction (HDR-3a):

$$(2)fdred[c] + hsdf[c] \xrightarrow{\text{MA2868/MA4236/MA4237}} com[c] + cob[c] + (2)fdox[c]$$

2) Electron bifurcation disulfide reduction:

Reaction:

$$2\ F_{420}H_2 + 2\ Ferredoxin + CoM - CoB \xrightarrow{\text{MA2868/MA4236/MA4237}} CoM + CoB + 2\ H_2\text{-} - Ferredoxin + 2\ F_{420}$$

Model Reaction (HDR-3b):

$$(2)f420 - 2h2[c] + (2)fdox[c] + hsdf[c] \xrightarrow{\text{MA2868/MA4236/MA4237}} com[c] + cob[c] + (2)fdred[c] + (2)f420 - 2[c]$$

MA0040 - Wyosine synthetase (New) This gene was neglected in prior reconstructions because it catalyzes a reaction that is not related to metabolism, rather with translation. In the interest of creating a complete model, especially one that would be amenable to acting as the base for a ME reconstruction it was added. It required adding the modified tRNA to the model (named trnaphe_imG14).

Reaction:

$$N1\text{-}Methylguanine\text{-}tRNA(Phe) + Pyruvate + S - Adenosyl\text{-}L\text{-}methionine \underset{MA0040}{\rightleftharpoons} 4\text{-}Demethylwyosine - tRNA(Phe) + L - Methionine + 5'\text{-}Deoxyadenosine + CO_2 + H_2O$$

Model Reaction (PHETRNAWY):

$$trnaphe[c] + amet[c] + pyr[c] \xrightleftharpoons{\text{MA0040}} trnaphe\_imG14[c] + met - L[c] + dad - 2[c] + co2[c] + h2o[c]$$

MA0117 - Dihydroneopterin aldolase, mptD (Modified) KEGG suggests has this gene annotated as the aldolase catalyzing the reaction named DHNPA2 and thus has been added as the GPR.

Reaction:

$$7,8\text{-}dihydroneopterin \xrightleftharpoons{\text{MA0117}} 6-(hydroxymethyl)\text{-}7,8-dihydropterin + glycolaldehyde + H^+$$

Model Reaction (DHNPA2):

$$dhnpt[c] \xrightleftharpoons{\text{MA0117}} 6hmhpt[c] + gcald[c] + h[c]$$

MA0696 - Dihydromethanopterin reductase (Mapped) Two papers have recently characterized a dihydromethanopterin reductase in methanogens that re catalyzes the reduction of 7,8-dihydromethanopterin to the tetrahydromethanopterin that is a precursor to the tetrahydrosarcinopterin that acts as a key cofactor in methanogenesis [410, 411]. The reducing equivalents to the enzyme, which was named DmrX [410], were not known, so we assumed they were provided by reduced coenzyme F420.

Reaction:

$$7,8\text{-}Dihydromethanopterin + F_{420}H_2 \xrightleftharpoons{\text{MA0696}} 5,6,7,8\text{-}tetrahydro-methanopterin + F_{420}$$

Model Reaction (H2MPTR):

$$h2mpt[c] + f420-2h2[c] \xrightleftharpoons{\text{MA0696}} h4mpt[c] + f420-2[c]$$

MA0246 - 3-octaprenyl-4-hydroxybenzoate carboxy-lyase (New) These genes likely plays a role either in lipid metabolism, or perhaps more interestingly, in the production of the polyprenyl tail of the cofactor methanophenazine. In absence of evidence for the later, the octaprenylphenol substrate was chosen. It required adding the 2-octaprenylphenol (named 2oppol) metabolite to the model.

Reaction:

$$2\text{-}octaprenylphenol + CO_2 \overset{\text{MA0246}}{\rightleftharpoons} 4\text{-}hydroxy\text{-}3\text{-}octaprenylbenzoate$$

Model Reaction (HBZDC):

$$2oppol[c] + co2[c] \overset{\text{MA0246}}{\rightleftharpoons} 3ophb[c]$$

MA0246 - 4-hydrozybenzoate decarboxylase This reaction is clear by homology:

Reaction:

$$4\text{-}hydroxybenzoate \overset{\text{MA0246}}{\rightleftharpoons} phenol + CO_2$$

Model Reaction (HBDC)

$$4hbz[c] \overset{\text{MA0246}}{\rightleftharpoons} phenol[c] + co2[c]$$

MA0711 - dimethylallyl-phosphate:FMNH2 prenyltransferase This reaction is involved in ubiquinone biosynthesis:

Reaction:

$$dimethylallyl\text{-}phosphate + reduced\ FMN \overset{\text{MA0711}}{\rightleftharpoons} prenylated\ FMN + orthophosphate$$

Model Reaction (HBDC)

$$dmp[c] + fmnh2[c] \overset{\text{MA0711}}{\rightleftharpoons} prfmnh2[c] + pi[c]$$

MA0694 - Triphosphoribosyl-dephospho-CoA synthetase (New) This gene catalyzes the transfer of triphosphate from ATP to dephospho-CoA. It necessitated adding the 2′-(5-Triphosphoribosyl)-3′-depospho-CoA metabolite to the model (named tpdpcoa).

Reaction:

$$ATP + Dephospho\text{-}CoA \overset{\text{MA0694}}{\rightleftharpoons} 2'\text{-}(5\text{-}Triphosphoribosyl)\text{-}3'\text{-}depospho\text{-}CoA + Adenine$$

Model Reaction (DPCOAPT):

$$atp[c] + dpcoa[c] \overset{\text{MA0694}}{\rightleftharpoons} tpdpcoa[c] + ade[c]$$

MA0696 - dihydromethanopterin reductase, dmrX (Modified) This gene was recently characterized as catalyzing a step in the biosynthetic pathway for the tetrahydrosarcinopterin cofactor in the homolog from *M. jannaschii*. This gene was added to the GPR of the reaction H2MPTR

Reaction:

$$F_{420}H_2 + dihydrosarcinopterin \overset{\text{MA0696}}{\rightleftharpoons} F_{420} + tetrahydrosarcinopterin +$$

H^+

Model Reaction (H2MPTR):

$$f420 - 2h2[c] + h2mpt[c] \overset{\text{MA0696}}{\rightleftharpoons} f420 - 2[c] + h4mpt[c] + h[c]$$

MA1068 - Glycosyl transferase family 2/dolichyl-phosphate beta-D-mannosyltransferase (Existing) The enzyme coded by this gene likely catalyzes one or both of two existing reactions that were missing annotations, either: 1) UDP-N-acetyl-D-glucosamine 2-epimerase or 2) Poly-galactosamine synthesis (cell wall). We assigned the gene to both reactions, due to this ambiguity. The resulting reactions are:

1) Reaction:

$$UDP\text{-}N\text{-}acetyl\text{-}D\text{-}glucosamine + H_2O \overset{\text{MA1068}}{\rightleftharpoons}$$
$$N\text{-}acetyl\text{-}D\text{-}mannosamine + UDP + H^+$$

Model Reaction (UAG2EMA):

$$h2o[c] + uacgam[c] \overset{\text{MA1068}}{\rightleftharpoons} h[c] + udp[c] + acmana[c]$$

2) Reaction:

$$UDP\text{-}N\text{-}acetyl\text{-}alpha\text{-}D\text{-}glucosamine \xrightarrow{\text{MA1068}}$$
$$Poly\text{-}N\text{-}acetylgalactosamine + H^+ + UDP$$

Model Reaction (PGAMS):

$$uacgam[c] \xrightarrow{\text{MA1068}} polyacgal[c] + h[c] + udp[c]$$

MA1068 - dolichyl-phosphate beta-D-mannosyltransferase This reaction is involved in lipid metabolism:

Reaction:

$$UDP\text{-}N\text{-}acetyl\text{-}D\text{-}glucosamine + H_2O \xrightarrow{\text{MA1068}} UDP + H^+ + N\text{-}acetyl\text{-}D\text{-}mannosamine$$

$$UDP\text{-}N\text{-}acetyl\text{-}D\text{-}glucosamine \xleftrightarrow{\text{MA1068}} Poly\text{-}N\text{-}acetylgalactosamine + UDP + H^+$$

Model Reaction (UAG2EMA/PGAMS)

$$uacgam[c] + h2o[c] \xrightarrow{\text{MA1068}} acmana[c] + udp[c] + h[c]$$

$$uacgam[c] \xleftrightarrow{\text{MA1068}} polyacgal[c] + h[c] + udp[c]$$

MA1170 - NAD(P)H hydrate hydro-lyase (New) This enzyme is involved in repairing hydrated nicotinamide nucleotide cofactors.

Reaction:

$$ADP + NAD(P)H - x \xrightarrow{\text{MA1170}} AMP + orthophosphate + NAD(P)H$$

$$NAD(P)H - x \xleftrightarrow{\text{MA1170}} (S) - NAD(P)H - x$$

Model Reaction (NNRDNADX/NNRDNADPX/NNRDRACENADH/ NNRDRACENADPH)

$$adp[c] + nad(p)hx[c] \xrightarrow{\text{MA1170}} amp[c] + pi[c] + nad(p)h$$

$$nad(p)hx[c] \xleftrightarrow{\text{MA1170}} S - nad(p)hx$$

MA1276 - Nucleotide Aminohydrolase (New) A homolog of this enzyme was characterized in *Streptomyces flocculus* [412].

Reaction:

$$S\text{-}methyl\text{-}5\prime\text{-}thioadenosine + H_2O + H^+ \xleftrightarrow{\text{MA1276}} S\text{-}methyl\text{-}5\prime\text{-}thioinosine + Ammonium$$

$$S\text{-}adenosyl\text{-}L\text{-}homocysteine + H_2O + H^+ \xleftrightarrow{\text{MA1276}} S\text{-}inosyl\text{-}L\text{-}homocysteine + Ammonium$$

Model Reaction (SMTANH/SATANH):

$$5mta[c] + h2o[c] + h[c] \xleftrightarrow{\text{MA1276}} 5mti[c] + nh4[c]$$

$$ahcys[c] + h2o[c] + h[c] \xleftrightarrow{\text{MA1276}} ihcys[c] + nh4[c]$$

MA1480 - Redox-dependent potassium symporter (New) Added a new reaction for a potassium-selective redox-dependent symport protein (Kt3r).

Reaction:

$$H_e^+ + K_e^+ \xleftrightarrow{\text{MA1480}} H^+ + K^+$$

Model Reaction (Kt3r):

$$h[e] + k[e] \xleftrightarrow{\text{MA1480}} h[c] + k[c]$$

MA1724 - Mechanosensitive Potassium Selective Channel (New) Added a reaction for mechanosensitive potassium chloride cotransport named KCmst.

Reaction:

$$K_e^+ + Cl_e^- \xleftrightarrows{\text{MA1724}} K^+ + Cl^-$$

Model Reaction (KCmst):

$$cl[e] + k[e] \xleftrightarrows{\text{MA1724}} cl[c] + k[c]$$

MA1816 - Serine-glyoxylate aminotransferase (New) The gene encodes a protein with annotations "Serine-pyruvate aminotransferase/archaeal aspartate aminotransferase". It could also be a glyoxylate aminotransferase.

Reaction:

$$L\text{-}serine + glyoxylate \xleftrightarrows{\text{MA1816}} 3\text{-}hydroxypyruvate + glycine$$

Model Reaction (SERPYRTA):

$$ser - L[c] + glx[c] \xleftrightarrows{\text{MA1816}} hpyr[c] + gly[c]$$

MA1816 - Serine-glyoxylate transaminase Based on homology, it appears that this gene encodes either a serine-glyoxylate transaminase or aspartate aminotransferase. As there exists the latter, the former was added as a reaction:

Reaction:

$$L\text{-}serine + glyoxylate \xrightarrow{\text{MA1816}} L\text{-}glycine + hydroxypyruvate$$

Model Reaction (SERTA)

$$ser\text{-}L[c] + glx[c] \xrightarrow{\text{MA1816}} hpyr[c] + gly[c]$$

MA2430 - 2-amino-4-hydroxy-6-hdroxymethyldihydropteride diphosphokinase, mptE (Modified) KEGG suggests has this gene annotated as the aldolase catalyzing the reaction named HPPK2 and thus has been added as the GPR.

Reaction:

$$ATP + 6\text{-}hydroxymethyl\text{-}7,8\text{-}dihydropterin \xleftrightharpoons{\text{MA2430}}$$
$$AMP + 6\text{-}hydroxymethyl\text{-}7,8\text{-}dihydropterindiphosphate + H^+$$

Model Reaction (HPPK2):

$$6hmhpt[c] + atp[c] \xleftrightharpoons{\text{MA2430}} 6hmhptpp[c] + amp[c] + h[c]$$

MA3306 - CDP-2,3-bis-(O-geranylgeranyl)-sn-glycerol synthase (Modified) The enzyme catalyzing the cytidylation of geranyl-geranyl-glycerol has recently been identified as MA3306 [413] rather than MA2010 which was inferred during initial model construction [284]. The associated gene has been changed from MA2010 to MA3306.

Reaction:

$$CTP + digeranylgeranylglyceryl\ phosphate + H^+ \xrightarrow{\text{MA3306}}$$
$$CDP\text{-}2,3\text{-}digeranylgeranyl\text{-}sn\text{-}glycerol + diphosphate$$

$$CTP + 3\text{-}hydroxy\text{-}digeranylgeranylglyceryl\ phosphate + H^+$$
$$\xrightarrow{\text{MA3306}} CDP\text{-}2,(3\prime\text{-}hydroxy)3\text{-}digeranylgeranyl\text{-}sn\text{-}glycerol +$$

diphosphate

Model Reaction (CDGGGS/2):

$$ctp[c] + dgggp[c] + h[c] \xrightarrow{\text{MA3306}} cdgggp[c] + ppi[c]$$

$$ctp[c] + 3hdgggp[c] + h[c] \xrightarrow{\text{MA3306}} 3hcdgggp[c] + ppi[c]$$

MA3554 - Cobalt ABC transport component, CbiM (Modified) MA3554 is a clear homolog to cbiM gene in known to be involved in cobalt transport.
Reaction:

$$Co_e^{2+} + ATP + H_2O \xleftrightarrow{\text{MA3554}} Co_c^{2+} + ADP + phosphate + H^+$$

Model Reaction (Coabc):

$$cobalt2[e] + atp[c] + h2o[c] \xleftrightarrow{\text{MA3554}} cobalt2[c] + adp[c] + pi[c] + h[c]$$

MA3706 - nucleoside-triphosphatase (Modified) It had been deomnstrated biochemically that MA3706 binds only ITP or XTP, thus it has been removed from NTP[1-9] [414].

MA3730 - Nucleotide phosphotransferase (New/Modified) MA3730 appears to be an adenylate kinase that may be associated with an existing reaction lacking a gene (ATP/GMP kinase, GK1) as well as a number of other phosphotransferase reactions.
Reaction:

$$ATP + GMP \xleftrightarrow{\text{MA3730}} ADP + GDP$$

$$ATP + AMP \xleftrightharpoons{\text{MA3730}} 2ADP$$

$$ATP + dAMP \xleftrightharpoons{\text{MA3730}} ADP + dADP$$

$$TTP + AMP \xleftrightharpoons{\text{MA3730}} ADP + TDP$$

Model Reaction (GK1/AK1/AK2/AK3):

$$atp[c] + gmp[c] \xleftrightharpoons{\text{MA3730}} adp[c] + gdp[c]$$

$$atp[c] + amp[c] \xleftrightharpoons{\text{MA3730}} (2)adp[c]$$

$$atp[c] + damp[c] \xleftrightharpoons{\text{MA3730}} adp[c] + dadp[c]$$

$$ttp[c] + amp[c] \xleftrightharpoons{\text{MA3730}} adp[c] + tdp[c]$$

MA3786 - Dihydroorotic acid dehydrogenase/(S)-dihydroorotate:fumarate oxidoreductase (New/Modified) Reaction:

$$(S)\text{-}dihydroorotate + Oxidized\ Coenzyme\ F_{420} + H^+ \xrightarrow{\text{MA0583/MA3786}} Orotate + Reduced\,Coenzyme\ F420$$

$$(S)\text{-}dihydroorotate + Fumarate \xrightarrow{\text{MA0583/MA3786}} Orotate + Succinate$$

Model Reaction (DHORD7/DHORDFUM)

$$dhor\text{-}S[c] + f420\text{-}2[c] + h[c] \xrightarrow{\text{MA0583/MA3786}} f420\text{-}2h2[c] + orot[c]$$

$$dhor\text{-}S[c] + fum[c] \xrightarrow{\text{MA0583/MA3786}} orot[c] + succ[c]$$

MA3990 - 8-oxo-dGTP diphosphohydrolase This reaction is involved in nucleotide metabolism:

Reaction:

$8\text{-}oxo\text{-}dGTP + H2O \xleftrightharpoons{\text{MA3990}} 8\text{-}oxo\text{-}dGMP + diphosphate$ Model Reaction (ODGTOPDP)

$8odgtp[c] + h2o[c] \xleftrightharpoons{\text{MA3990}} 8ogmp[c] + ppi[c]$

MA3981/MA3108 - Glutathione-Regulated Potassium Efflux (New) Added a reaction for potassium regulated efflux (Kgtht).

Reaction:

$$K^+ + H_e^+ \xrightarrow{\text{MA3981/MA3108}} K_e^+ + H^+$$

Model Reaction (Kgtht):

$$k[c] + h[e] \xrightarrow{\text{MA3981/MA3108}} k[e] + h[c]$$

MA3994 - 4-[N-γ-L-glutamyl-)-p-(β-aminoethyl)phenoxy-methyl]-2-(aminomethyl)furan glutamate extension (Modified) The gene MA3994 has long been recognized as a marker of methanogenesis (InterPro: putative methanogenesis marker protein 15). We hypothesize that the enzyme plays the role in methanofuran biosynthesis. The enzyme catalyzing the final steps of methanofuran biosynthesis, which consist of somewhere between 6 and 11 glutamate extension event, has so far been unidentified [204]. The enzyme contains a glutamate binding domain (MutL: glutamate mutase). Additionally, it contains conserved Hydantoinase/oxoprolinase N-terminal domain that is found in only one characterized protein, mfnF, the gene

catalyzing the previous step in methanofuran biosynthesis [204]. As such, we suggest MA3994 be named mfnG and catalyze several reactions.

Reaction:

$$p\text{-}(\beta\text{-}aminoethyl)phenoxy\text{-}methyl\text{-}2\text{-}(aminomethyl)furan + L\text{-}Glu + ATP \xrightarrow{\text{MA3994}}$$

$$4\text{-}[N\text{-}\gamma\text{-}L\text{-}glutamyl\text{-})\text{-}p\text{-}(\beta\text{-}aminoethyl)phenoxy\text{-}methyl]\text{-}\ 2\text{-}(aminomethyl)furan + ADP + orthophosphate + H_2O$$

Model Reaction (MFRS9/MFRS10/MFRS11/MFRS12)

$$aepmamfr[c]/4gaepmamfr[c]/4ggaepmamfr[c]/4gggaepmamfr[c] + glu-L[c] + atp[c] \xrightarrow{\text{MA3994}} 4gaepmamfr[c]/4ggaepmamfr[c]/4gggaepmamfr[c]/mfr - b[c] + h2o[c] + adp[c] + pi[c]$$

MA4095 - 2,5-diamino-6-ribitylamino-4(3H)-pyrimidinone 5′-phosphate deaminase (MAPPED) MA4095 encodes a putative pyrimidine deaminase and is also in a highly conserved gene cluster with MA4092. MA4092 is known to catalyze the reaction DROPPRx/y. The next reaction downstream is a deamination reaction, therefore we hypothesize MA4095 is the enzyme:

Reaction:

$$2,5\text{-}diamino\text{-}6\text{-}ribitylamino\text{-}4(3H)\text{-}pyrimidinone5'\text{-}phosphate + H_2O + H^+ \xrightarrow{\text{MA4095}} 5\text{-}Amino\text{-}6\text{-}(5'\text{-}phosphoribitylamino)uracil + NH_4$$

Model Reaction (DRTPPD)

$$25dthpp[c] + h[c] + h2o[c] \xrightarrow{\text{MA4095}} 5aprbu[c] + nh4[c]$$

MA4330 - GDPmannose:dolichyl-phosphate O-beta-D-mannosyltransferase (NEW) This reaction is involved in lipid metabolism:

Reaction:

$$GDP\text{-}D\text{-}mannose + Dolichyl\text{-}phosphate \xleftrightharpoons{\text{MA4330}} GDP + Dolichyl\ phosphate\ D\text{-}mannose$$

Model Reaction (GMANDOLP)

$$gdpmann[c] + dolichylp[c] \xleftrightharpoons{\text{MA4330}} gdp[c] + dolichylpman[c]$$

MA4337 - Ferrous iron ABC transport component, FeoB (Modified) MA4338 is a clear homolog to feoB gene in *E. coli* which is known to transport ferrous iron.

Reaction:

$$Fe_e^{2+} + ATP + H_2O \xleftrightharpoons{\text{MA4337/MA3478/MA3479}} Fe_c^{2+} + ADP + phosphate + H^+$$

Model Reaction (FE2abc):

$$fe2[e] + atp[c] + h2o[c] \xleftrightharpoons{\text{MA4337/MA3478/MA3479}} fe2[c] + adp[c] + pi[c] + h[c]$$

MA4504 - D-arabino-hex-3-ulose-6-phosphate isomerase This reaction is involved in central lipid metabolism:

Reaction:

$Formaldehyde + D\text{-}Ribulose\ 5\text{-}phosphate \xleftrightharpoons{\text{MA4504}} D\text{-}arabino\text{-}Hex\text{-}3\text{-}ulose\ 6\text{-}phosphate$

$D\text{-}arabino\text{-}Hex\text{-}3\text{-}ulose\ 6\text{-}phosphate \xleftrightharpoons{\text{MA4504}} D\text{-}Fructose\ 6\text{-}phosphate$

Model Reaction (UAG2EMA/PGAMS)

$fald[c] + ru5p\text{-}D[c] \xleftrightharpoons{\text{MA4504}} ah6p\text{-}D[c]$

$ah6p\text{-}D[c] \xleftrightharpoons{\text{MA4504}} f6p[c]$

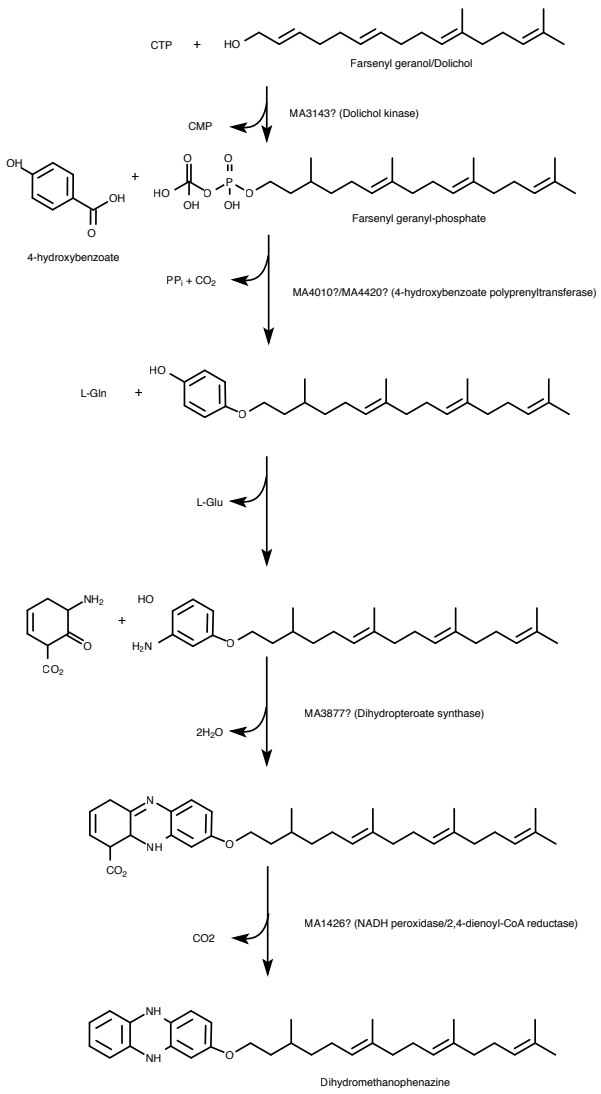


Figure 5.11: **Hypothesized Methanophenazine Biosynthesis Pathway.** Each step is a hypothesis based on putative function of the indicated gene product. Locus tags from *M. acetivorans* C2A are shown.

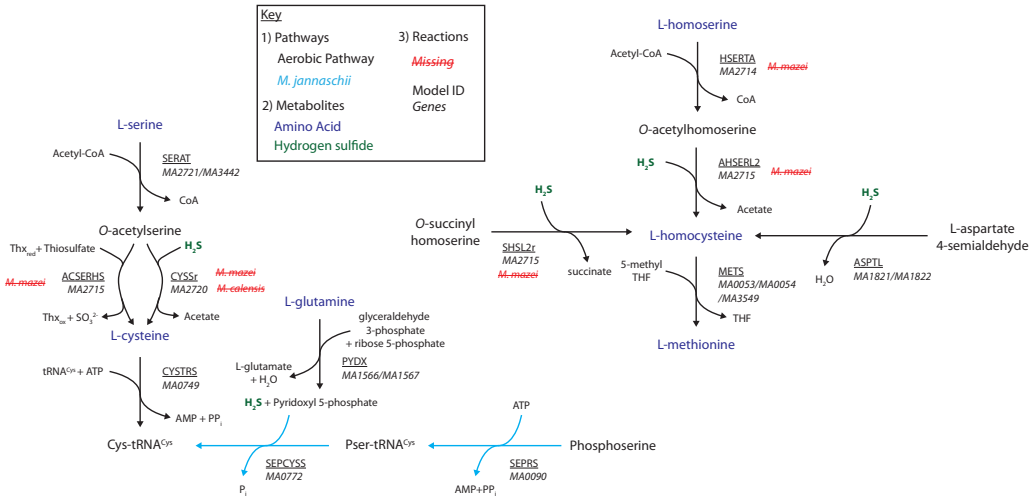


Figure 5.12: **Cysteine & Methionine Biosynthesis Pathways** Several pathways for biosynthesis of cysteine and methionine exist in many of the *Methanosarcina* spp; however, several are missing within the *M. mazei* and *M. calensis* groups. For example, the ancestral methanogen pathway to generate cysteine from phosphoserine must be employed by *M. mazei* while all the other *Methanosarcina* can use the pathway found in aerobic methanogens. Similarly, the *M. mazei* are missing the genes needed to generate L-homocysteine, a precursor for L-methionine, from L-homoserine, requiring it to be generated from L-aspartate 4-semialdehyde instead. Genes locus tags are those from *M. acetivorans* C2A. Reaction names are those used in the metabolic models and stand for: **ACSERHS** – *O*-acetyl-L-serine:hydrogen-sulfide 2-amino-2-carboxyethyltransferase **AHSERL2** – *O*-acetylhomoserine (thiol)-lyase, **ASPTL** – semialdehyde thiolase, **CYSSr** – cysteine synthase, **CYSTRS** – cysteinyl-tRNA synthetase, **HSERTA** – homoserine O-trans-acetylase, **METS** – methionine synthase, **PYDX** – pyridoxal 5-phosphate synthase, **SEPCYSS** – *O*-phospho-L-seryl-tRNA:Cys-tRNA synthase, **SEPRS** – *O*-phospho-L-serine:tRNA(Cys) ligase (AMP-forming), **SERAT** – serine *O*-acetyltransferase, **SHSL2r** – *O*-succinyl-L-homoserine succinate-lyase (adding hydrogen sulfide),

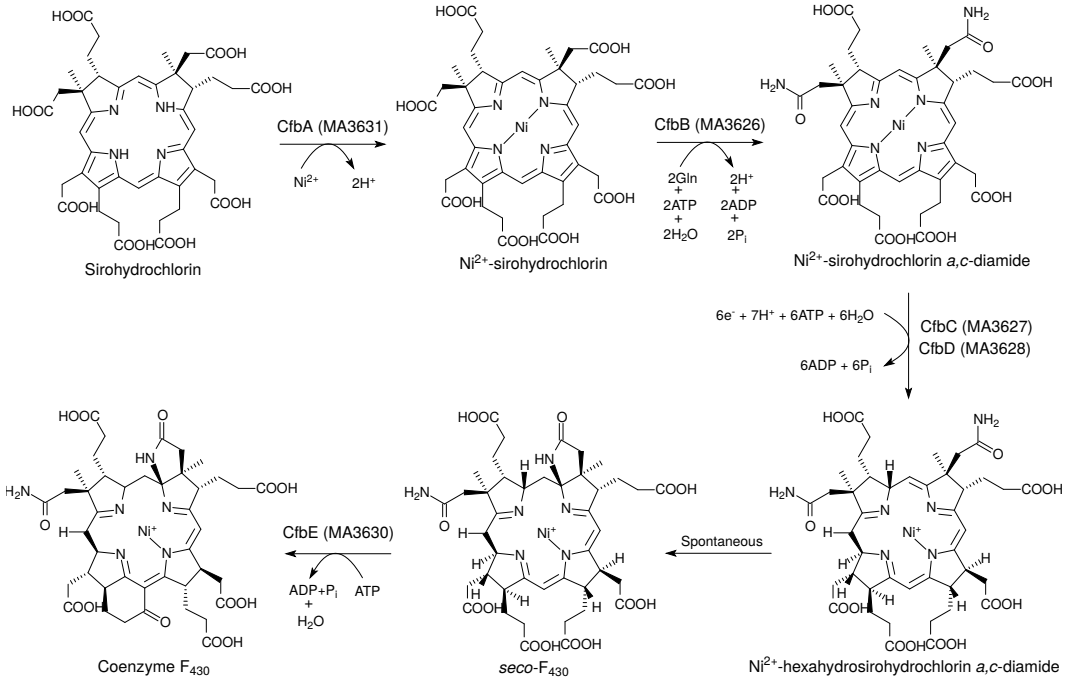


Figure 5.13: **Coenzyme F430 Biosynthesis Pathway.**

Part II

Stochastic and Continuum Analyses of Heterogeneity in *Escherichia coli*

Chapter 6

Effects of DNA Replication on mRNA Noise

There are several sources of fluctuations in gene expression. Here we study the effects of time-dependent DNA replication, itself a tightly controlled process, on noise in mRNA levels. Stochastic simulations of constitutive and regulated gene expression are used to analyze the time averaged mean and and variation in each case. The simulations demonstrate that in order to capture mRNA distributions correctly, chromosome replication must be realistically modelled. Slow relaxation of mRNA from the low copy number steady-state prior to gene replication to the high steady-state after replication is set by the transcript's half-life and contributes significantly to the shape of the mRNA distribution. Consequently both the intrinsic kinetics and the gene location play an important role in accounting for the mRNA average and variance. Exact analytic expressions for moments of the mRNA distributions that depend on the DNA copy number, gene location, cell doubling time, and the rates of transcription and degradation are derived for the case of constitutive expression and subsequently extended to provide

The contents of this chapter are based in part on work previously published as Joseph R. Peterson John A. Cole, Jingyi Fei, Taekjip Ha and Zaida Luthey-Schulten. "Effects of DNA Replication on mRNA Noise," *Proceedings of the National Academy of Sciences of the USA* 112(52):15886 (2015) [50] and Tyler M. Earnest, John A. Cole, Joseph R. Peterson, Thomas E. Kuhlman, and Zaida Luthey-Schulten. "Ribosome biogenesis in replicating cells: integration of experiment and theory," *Biopolymers* 105(10):735 (2016) [54]. The contribution by J.A.C. was indispensable providing nearly all of the derivations and created figure 6.20. J.F. performed the smFISH experiments measuring the mRNA counts of *ptsG* in *E. coli* and also created figure 6.6.

approximate corrections for regulated expression and RNA polymerase variability. Comparisons of the simulated models and analytical expressions to experimentally measured mRNA distributions show that they better capture the physics of the system than previous theories.

6.1 Introduction

Every step in the process of gene expression includes some inherent randomness. This may stem from the intrinsically stochastic nature of chemical reactions, chance differences between cells in the numbers of available reactants, intracellular crowding, or any of a number of other sources of biological variability [40, 43, 415–417]. All told, noisy gene expression has profound effects on cellular behaviour at both the individual and population levels, enabling switching between phenotypes by individual cells [41, 46, 418–421] as well as the potential for entire populations to divergently adapt to multiple niches within their environment [422]. As a result, a great deal of work over the last decade has focused on understanding and quantifying the various sources of biological stochasticity.

In a series of now-classic papers, theorists and experimentalists alike have shown that the equations governing stochastic gene expression elicit steady-state distributions of proteins and mRNA in good agreement with observations [38, 41, 42, 418, 423–425]. Many of these works have also considered forms of transcriptional regulation wherein a gene can switch between active and inactive transcriptional states (either through the binding of a

transcription factor [42, 419, 422, 426, 427], or through structural changes to the DNA that may occlude transcription start sites [45, 428]). More recently, researchers have begun to venture beyond the steady-state approximation in order to address sources of noise that are tied to cell cycle-dependent processes. By considering mixtures of steady state mRNA distributions associated with one and two copies of the DNA, Jones *et al.* [49] was the first to show that the duplication of a gene during replication can directly contribute to the observable noise in mRNA copy number. They used these results to partition experimentally observed mRNA noise into contributions associated with gene duplication, variability in RNA polymerase copy numbers, and experimental error [49].

In this paper we perform stochastic simulations, exactly sampling chemical master equations (CME) that explicitly account for chromosome replication, in order to show that gene duplication does in fact contribute to the observed variations in mRNA levels. We find, however, that our simulated results differ consistently and often significantly from the predictions of Jones *et al.* We show that after gene duplication, a cell's mRNA count relaxes slowly from a low state (associated with the initial gene copy number) to a high state (associated with the copy number after replication) at a rate proportional to the mRNA half-life, a transition that can take several minutes and account for a significant portion of the overall cell cycle (see Fig. 6.1). This seemingly minor effect can lead to divergence between the predicted and simulated mRNA Fano factors (a measure of the "noisiness" of the transcribed mRNA and equal to the variance over the mean) of 20% to greater

than 80%, depending on the cell doubling time, the location of the gene on the chromosome, and the mRNA degradation rate. Such errors can easily lead to misattribution of observed mRNA variability to spurious sources, and cloud the interpretation of experimental results.

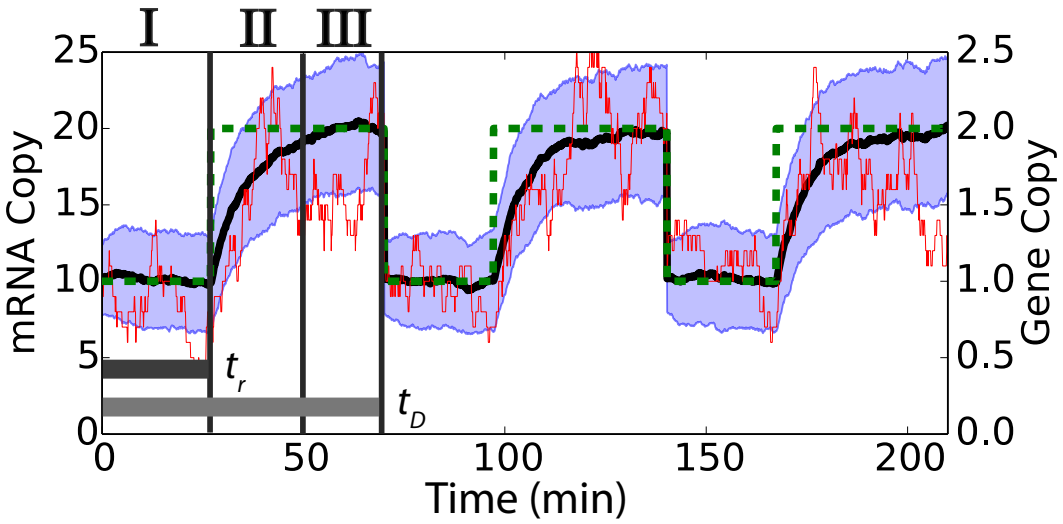


Figure 6.1: **Simulation Schematic.** A schematic composed of 200 simulation replicates showing the progress of the average mRNA count (black line) before and after a gene duplication event (traced by green dotted line). The area encompassing the average $\pm 1\sigma$ (blue) are shown along with an example simulation trace (red). Gene duplication is followed by a transient period where the mRNA relaxes from an initially low to a high count at a rate proportional to the degradation rate of the mRNA. Three regions exist and are delineated by vertical lines: A pre-duplication state (I) wherein the mRNA is in a low copy number steady-state, a relaxation period just after duplication (II) where the mRNA relaxes up to a new equilibrium steady-state (III). In these simulations the doubling time (t_D) was taken to be 70 minutes, the total DNA replication time was taken to be 45 minutes, the gene was positioned 55% of the way from the origin to the terminus ($t_r \approx$27 minutes), the transcription rate k_t was 1.26 molecules/min and the degradation rate k_d was 0.126/min.

Our findings motivated a time-dependent analytical treatment of the

noise contribution originating from gene duplication, as well as several corrections in order to account for transcriptional regulation and variability in RNA polymerase (RNAP) and transcription factor copy numbers. The expressions are nearly exact for the case of constitutive transcription, even when including RNAP noise, and show extremely good agreement with both simulations and experiments when accounting for regulation. These results demonstrate that the explicit treatment of gene replication and careful accounting for the subsequent product copy number relaxation time is necessary for accurately describing mRNA—and in turn protein—variability.

6.2 Methods

Simulations were performed using the Gillespie stochastic simulation algorithm [76] as implemented in the Lattice Microbes software version 2.2 [77, 429]. All simulations were performed using NVIDIA GPUs and analysis was written in Python using the PyLM interface to Lattice Microbes version 1.0 [78]. Input files can found with the Supplemental Files.

Both a constitutive model of gene expression and a two-state model of gene expression were considered (Eqs. 6.1 and 6.6). Doubling times of 40 or 70 minutes were examined and cell division was implemented by dividing the gene counts in half and binomially distributing the mRNA count between the cells with equal probability. The replication time (t_r) as well as the number of genes and replication forks at the start of the cell cycle are based on the theory of Cooper and Helmstetter [430]. The DNA replication

time was taken as 45 minutes; a value close to the average measured [431]. When simulating regulation, the gene states were randomized at division time with probability to be active $P_{\text{on}} = k_{\text{on}}/\left(k_{\text{on}} + k_{\text{off}}\right)$. The transcription rate constants k_t and k_d, were varied as described in the main text. For each set of rate parameters, three technical replicate simulations were run each of which included independent trajectories of 200 cell lineages growing for 10 generations.

6.3 Results

6.3.1 Explicit Simulation of Gene Duplication for Constitutive Expression

Stochastic simulations of gene expression were used to determine the effect of chromosome replication on mRNA noise. A constitutive model of gene expression (Eq. 6.1) wherein mRNA is transcribed from its gene at rate k_t and degraded at rate k_d, is considered first, as the majority of genes are under no regulatory control under physiological conditions.

$$\begin{aligned} D &\xrightarrow{k_t} D + m \\ m &\xrightarrow{k_d} \emptyset \end{aligned} \tag{6.1}$$

Simulations were performed using Gillespie's stochastic simulation algorithm (SSA) [76] as implemented in our Lattice Microbe software [77]. For each simulation replicate, the mRNA copy number was tracked within a

single lineage spanning 10 full cell cycles. Starting from a defined initial state, the copy number was allowed to evolve until t_r (the replication time for the gene) at which time the gene copy number was doubled to model the effect of replication. The simulations then continued until t_D (the division time) at which time the intracellular components were halved to account for cell division, and the next generation in the lineage begins. An example of this process is shown in Fig. 6.1. The cell cycle length, replication time, and transcription rate were all varied, but the mRNA degradation rate was held fixed at 0.126 min^{-1} in order to maintain the average mRNA half-life in *E. coli* of 5.5 min [159]. Two different cell doubling times were studied—70 minutes and 40 minutes. Genome replication in *E. coli* requires ~ 45 minutes and is relatively insensitive to changes in growth rate or culturing conditions [430–432]. As such, cells doubling in less than this amount of time must maintain multiple chromosome replication forks at different stages of completion. This means that depending on their location along the genome, some genes in our fast-growing cells ($t_D = 40$ minutes) exist with either 2 or 4 copies (we ignore the short-lived 3-copy state that arises when one replication fork briefly out-paces the other [433]) while others exist with either 1 or 2 copies (see Fig. S6.5B and SI Table S6.8C). In the slow-growing cells ($t_D = 70$ minutes) there exist either 1 or 2 copies of all genes. For both doubling times, simulations were performed across a series of replication times (t_r) corresponding to genes located across the genome (spanning from the origin of replication to the replication terminus in 5% increments).

Our simulations show that after gene duplication the mean mRNA count

relaxes to twice its prior value on a time-scale that is set by the mRNA's half-life. This, it turns out, can constitute a significant portion of the cell cycle (in *E. coli* the average mRNA half-life is $\sim 14\%$ of a 40 minute cell cycle, and the total relaxation takes about 40% of the cell cycle) and significantly impact the statistics of observable mRNA copy numbers (see Fig. S6.6).

6.3.2 Analytical Time-Dependent mRNA Statistics for Constitutive Expression

We derived expressions for the mean, variance, and Fano factor of an mRNA being constitutively expressed from a gene that is duplicated during the cell cycle (a detailed description can be found in the Supporting Information Section 1.1). This work hinged on the fact that the mean mRNA copy number, $\langle m \rangle$, over an ensemble of cells can be written as a time-average over the instantaneous mean copy number, $\bar{m}(t)$, and likewise, the variance, Var$[m]$, can be written in terms of a time-average over the instantaneous variance, $\sigma_m^2(t)$, and the square of the mean copy number (see SI Eqs. S6.31 & S6.33). Differential equations for the instantaneous mean and variance were derived from the chemical master equation (see SI Eqs. S6.9–6.28), and solved to yield:

$$\sigma_m^2(t) = \bar{m}(t) = \begin{cases} \frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{6.2}$$

Interestingly, the mRNA remains Poisson-distributed after gene duplication

as it relaxes to its new steady state (*i.e.* the probability of measuring m mRNA in a cell at an amount of time t after the start of its cell cycle can be written $P(m|t) = \text{Pois}(\bar{m}(t))$). Time-averaging over the cell cycle yielded the expressions:

$$\langle m \rangle = \langle m \rangle_1 \left[1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D} \right] \tag{6.3}$$

$$\begin{aligned} \text{Var}\,[m] &= \langle m \rangle - \langle m \rangle^2 \\ &\quad + \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2 k_d t_D} \right] \end{aligned} \tag{6.4}$$

$$\begin{aligned} \text{Fano}\,[m] &= 1 - \langle m \rangle \\ &\quad + \frac{\langle m \rangle_1^2}{\langle m \rangle} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2 k_d t_D} \right] \end{aligned} \tag{6.5}$$

where $\langle m \rangle_1 = k_t / k_d$ represents the mean mRNA copy number prior to gene duplication, and $f = (t_D - t_r)/t_D$ represents the fraction of the cell cycle after the gene duplication event. Although these results were derived assuming that the ages of cells in a population should be uniformly distributed, log-phase populations are in fact known to have exponentially distributed ages [432, 434]. This can be easily accounted for analytically (see Equations S6.30–6.37), but it amounts to a fairly small correction ($< 10\%$, see Figure S6.7), and significantly complicates the expressions. It is worth noting that in the limit where the mRNA degradation rate, k_d, becomes large, relaxation after gene duplication becomes instantaneous, and our "time-dependent" (TD) theory reduces to the "time-independent" (TI) theory of Jones *et al.* [49]. In the limit

of slow mRNA degradation, the mRNA distribution never relaxes to the high state, and cells remain in the low copy number state until division.

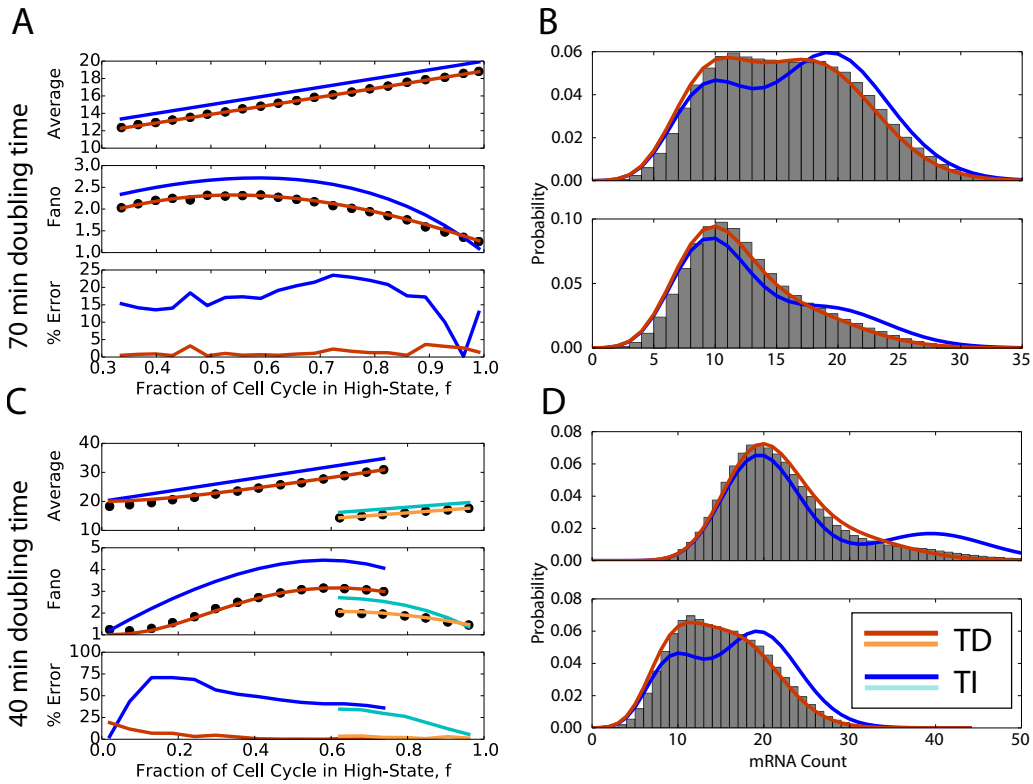


Figure 6.2: **Time-Dependent and Time-Independent Theories.** A) The mean, Fano factor, and relative error in the Fano factor for slow growing cells (70 min doubling time). Black dots represent the results of 200 simulated replicates, while orange and blue lines represent the TD and TI theories, respectively. B) Comparison of mRNA distributions for slow growing cells. The grey histogram represents the results of simulations, while the orange and blue lines again represent the TD and TI theories. The top distribution is that of a gene copied half-way through the cell cycle ($f = 0.5$) while the bottom distribution is that of a gene copied at the biginning of the cell cycle ($f \approx 1.0$). C & D) Statistics (mean, Fano factor, and relative error) and distributions for fast growing cells (40 min doubling time). Genes can exist in either 2 or 4 copies (deep orange and blue) or 1 or 2 copies (light orange and blue), depending on their location along the chromosome. In all cases, the time-dependent theory better captures simulation data.

Comparison of Eqs. 6.3 and 6.5 with simulations demonstrates the accuracy of the time-dependent theory (see Figs. 6.2A & C, S6.9, and S6.8). For both doubling times our expressions for $\langle m \rangle$ and the Fano factor prove nearly exact, whereas the time-independent theory tends to overestimate both values. Comparing the shape of the mRNA distributions proves equally impressive. Numerically time-averaging $P(m|t)$ yields distributions that strongly agree with histograms of our simulated mRNA counts (see Figs. 6.2B & D, S6.10, and S6.11, orange lines). In order to quantify the agreement of the time-dependent and time-independent models, we computed the Kullback-Leibler divergence (Eq. S6.76) between simulated and theoretical distributions. The divergence from our simulated distributions is approximately 10-fold smaller when using the time-dependent theory, but we note that this improvement breaks down within a narrow range of gene loci in the fast-growing cells (see Fig. S6.12). This disagreement occurs among genes located between about 50 and 70% of the way from the origin to terminus, and is due to the fact that these genes are duplicated very late in the 40 minute cell cycle. The associated mRNA counts has insufficient time to relax to their post-duplication steady-states, and upon division, they drop well below their pre-duplication steady-state (see Fig. S6.13A). As a result, the dynamics of these mRNA are better modelled assuming both early and late relaxations (see SI Section 1.2, Eqs. 6.45-6.46, and Fig. S6.13B–C).

Extending our comparisons to experimental data proves similarly fruitful. Theoretical distributions computed using measured k_d, f and t_D were compared to 26 previously reported experimental data sets [49,156]. A few

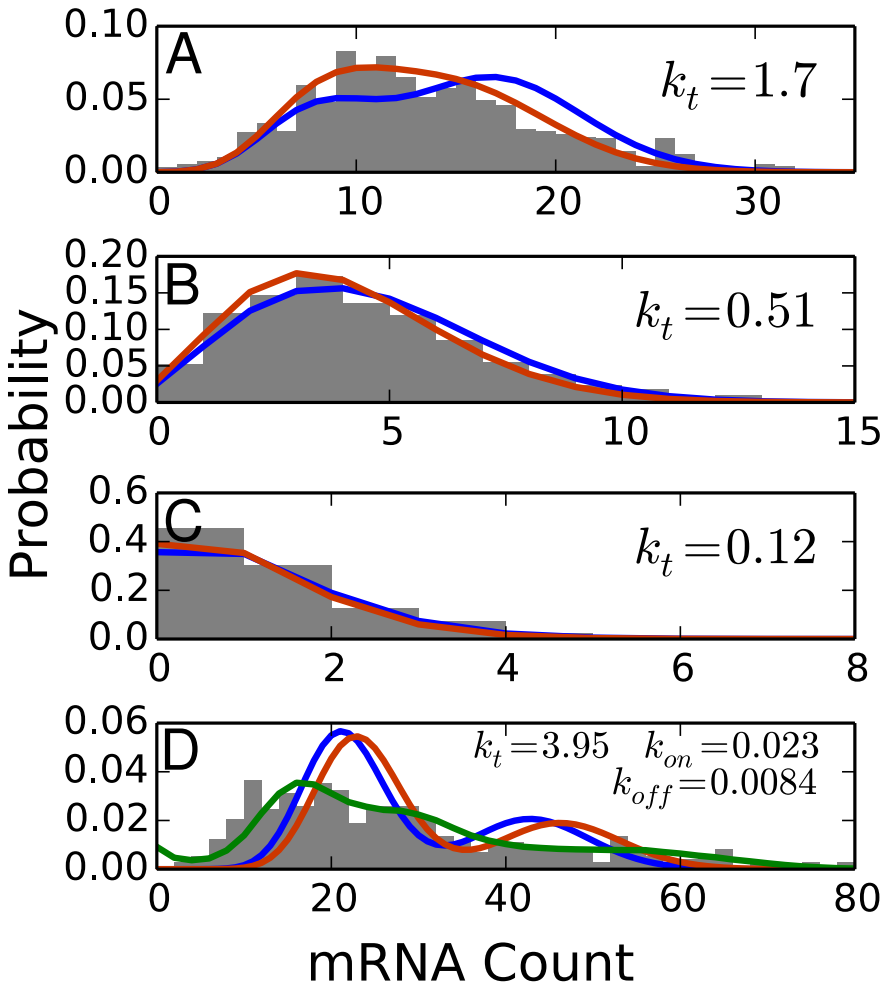


Figure 6.3: **Comparison to Experiments.** A comparison of predicted distrubtions computed assuming constitutively expressed genes with (orange lines) and without (blue lines) accounting for the time-dependence of the mRNA relaxation to experimental data for A-C) various *lac* promoter mutants [49] and D) *ptsG* [156]. The *lac* mRNA has a half–life of 5.5 min and spends 2/3 of the cell cycle after gene replication, while the *ptsG* transcript has a half-life of 2.8 min and spends about 1/3 of the cell cycle after gene replication. A fit to the regulated model shows much better agreement for *ptsG* (green line). All rates from the fits are given in units of per minute. See SI Section 1.9 and Figs. S6.14, S6.15, and S6.16 for further comparisons and details.

representative distributions for genes with different values of f, k_d, and $\langle m \rangle$ are shown in Fig. 6.3. The time-dependent theory outperforms the time-independent theory in all cases, clearly demonstrating its utility (see SI Section 1.9 and Figs. S6.14 and S6.15). We note, however, that neither theory performs well when fitting mRNA distributions for strongly regulated genes. Figure 6.3D shows the distribution of *ptsG* mRNA counts in single *E. coli* cells obtained via super-resolution imaging and modeling [156]. This gene is known to be regulated via transcription factors and small RNA [156, 435]). The orange and blue lines show theoretical distributions computed according to the time-dependent and time-independent treatments. Both curves are underdispersed, indicating the need for a model that directly accounts for transcriptional regulation (green line; discussed in the next section and SI Section 1.9; also see Figs. S6.16 and S6.17).

6.3.3 Corrections to the Analytical Model for Regulation, as Well as RNAP and TF Variability

Several corrections to the our time-dependent analytical model were derived in order to account for other sources of noise. The first and most important is a correction that approximates the noise stemming from transcriptional regulation (see SI Section 1.3). A gene is modelled as being in either an "off" or "on" state; the "on" state is capable of producing mRNA at rate k_t, while the "off" state is silenced. Genes are allowed to switch between states at rates k_{on} and k_{off}; schematically:

$$
\begin{aligned}
D_{\text{off}} &\underset{k_{\text{off}}}{\overset{k_{\text{on}}}{\rightleftharpoons}} D_{\text{on}} \\
D_{\text{on}} &\xrightarrow{k_t} D_{\text{on}} + m \\
m &\xrightarrow{k_d} \emptyset
\end{aligned}
\tag{6.6}
$$

In the limit where the transcriptional state switching is fast compared to the mRNA degradation rate (*i.e.* $k_{\text{on}}, k_{\text{off}} >> k_d$, meaning the average number of "on" genes relaxes quickly to its new steady-state after gene replication) and k_{on} is greater or at least of similar order as k_{off}, the Fano factor can be approximated with the addition of a single term:

$$
\begin{aligned}
\text{Fano}\,[m] \approx 1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} - \langle m \rangle \\
+ \frac{\langle m \rangle_1^2}{\langle m \rangle} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2 k_d t_D} \right]
\end{aligned}
\tag{6.7}
$$

where, again, $\langle m \rangle$ is given by Eq. 6.3, but now $\langle m \rangle_1 = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \frac{k_t}{k_d}$. It is important to note that this result is based on the assumption that the relaxation of the instantaneous mRNA mean ($\bar{m}(t)$) and variance ($\sigma_m^2(t)$) occur on a similar timescale. Our own simulations indicate that this approximation may not in general be true, but it drastically simplifies the analysis and keeps the resulting expressions for the mean, variance, and Fano factor tractable. Within appropriate parameter ranges, we find good agreement with simulation (see Fig. S6.18), but we note that when the gene switching rates are slow, or significantly favor the "off" state (meaning the mRNA is especially "bursty") Eq. 6.7 shows poorer agreement. As a result, further corrections were derived for cases in which $k_{\text{on}}, k_{\text{off}} \lesssim k_d$ (see Supporting Information

Section 1.4). This refined analysis treats the mean number of "on" genes as a dynamic variable after gene duplication (rather than assuming rapid relaxation) and yields somewhat unwieldy expressions for $\langle m \rangle$ and Var$[m]$ which themselves depend on whether the regulation is controlled by a repressor- or activator-type transcription factor (see Supporting Information Eqs. S6.62, S6.63, S6.64, and S6.65).

Because gene transcription depends on the activity of a number of proteins including RNA polymerase (RNAP) and any of several transcription factors (TFs), variability in these proteins' copy numbers can naturally impact mRNA levels within the cell. We considered how our time-dependent theory's results change when the numbers of either RNAP or an activator-type TF were assumed to vary (leading to variation in the effective transcription and gene activation rates, see Supporting Information Sections 1.5 & 1.6).

The Fano factor correction derived for RNAP-associated noise resulted in a simple additive term:

$$\begin{aligned} \text{Fano}[m] \approx\ & \text{Fano}[m]_{\bar{k_t}} \\ & + \frac{\langle m \rangle_{1,\bar{k_t}}^2}{\langle m \rangle_{\bar{k_t}}} \frac{\text{Var}[k_t]}{\bar{k_t}^2} \left[1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D} \right] \end{aligned} \tag{6.8}$$

where $\bar{k_t}$ represents the mean value of k_t, $\langle m \rangle_{\bar{k_t}}$ and Fano$[m]_{\bar{k_t}}$ are the mean mRNA number and Fano factor evaluated according to Eqs. 6.3 and 6.5 assuming $k_t = \bar{k_t}$, and $\langle m \rangle_{1,\bar{k_t}} = \bar{k_t}/k_d$ is the mean mRNA count prior to gene duplication. If the RNAP copy numbers are Γ-distributed, then Var$[k_t] \approx$

$k_{t,0}^2 \frac{\langle R \rangle}{\beta}$ where β represents the "rate" parameter of the distribution. For an *E. coli* doubling in approximately 40 min, the mean RNAP copy number has been measured to be $\sim 3,000$ per cell [436], placing it well into the "extrinsic noise limit" (for which $\sigma^2/\mu^2 \approx 0.1$ [424]) implying that β can be approximated as $1/300$. Inserting this into Eq. 6.8, we find that the contribution to the Fano factor from RNAP copy number variability can be roughly approximated as $\langle m \rangle / 10$, in accordance with [49] (see Fig. S6.19).

In contrast, the correction derived for TF-associated noise resulted in a cumbersome expression, which, when evaluated across a range of k_{off} and $\bar{k_{\text{on}}}$ values (where $\bar{k_{\text{on}}}$ represents the mean value of k_{on}), tended to be relatively small. We found it approached $\langle m \rangle / 10$ only when $k_{\text{off}} >> \bar{k_{\text{on}}}$, and in cases where $k_{\text{off}} \lesssim \bar{k_{\text{on}}}$ we found this correction remained well below $\sim 3\%$ of $\langle m \rangle$. Importantly, these results indicate that TF-associated variability generally imparts less mRNA noise than does RNAP-associated variability.

The corrections for RNAP and TF-associated noise resulted from the promotion of certain rates—k_t and k_{on}, respectively—to random variables and Taylor expanding about their means. Similar analyses can be performed for other potential sources of noise, including variability in t_r or t_D; in both cases, however, experiments show that the variance of these parameters is generally much less than 10% of their mean [433], and thus they are not likely to significantly impact measurable mRNA noise.

The analytic expressions derived here can be leveraged to greatly simplify the determination of kinetic parameters. The fitting of the *ptsG* mRNA distribution in Fig. 6.3D exemplifies this; it cannot be fit without accounting

for transcriptional regulation but this requires the simultaneous varying of k_{on}, k_{off}, and k_t. Equations 6.5, 6.7, and 6.8 can be used to solve for k_{on} and k_{off} as functions of k_t (see SI section 1.9), meaning that the fitting problem can be reduced to a simple 1-D scan over possible values for the transcription rate (assuming fixed k_d and $\langle m \rangle$). This significantly simpler problem was then performed numerically and resulted in the values $k_t = 3.95$ min^{-1}, $k_{\text{on}} = 0.023$ min^{-1}, and $k_{\text{off}} = 0.0084$ min^{-1}, all of which are physiologically reasonable [437].

6.4 Discussion

We have computationally and analytically studied the effects of DNA replication on mRNA noise. By formulating the process in terms of a CME, we were able to determine the time-dependent mean, variance, and distribution of the mRNA as a function of its degradation and transcription rates, the cell's doubling time, and the gene's position on the chromosome. We have found that failure to account for the slow relaxation of the messenger distribution to its post-duplication steady-state results in overestimation of the associated noise. Importantly, this overestimation can have a profound impact on the interpretation of both experimental and theoretical results.

As a hypothetical example, consider a single cell mRNA counting experiment in which the cell doubling time is measured to be 40 min, the mean count to be 10 messengers per cell each with the average degradation rate in *E. coli* of 0.126 min^{-1}, with the gene of interest being located roughly

one-third of the way between the origin and terminus of replication. Assume the Fano factor for the population was measured to be 3.25. These reasonable values can lead to very different interpretations of experiential data depending on how gene-duplication is treated. Prior to the study by Jones *et al.* [49], the entirety of the noise larger than 1 might have been attributed to transcriptional regulation and extrinsic factors like RNAP variability. In that case, after accounting for the RNAP noise contribution, it would have been concluded that the gene was quite strongly regulated (see Fig. S6.20, left bar). After [49], exactly the opposite conclusion could have been reached—essentially all of the observed noise could be attributed to RNAP variability and gene duplication. It would have appeared that there was no evidence of transcriptional regulation (see Fig. S6.20, middle bar). In fact, our analysis shows that both gene duplication and regulation contribute similar but modest amounts to the overall noise level (see Fig. S6.20, right bar). We note that this example is a special case, and in general the different models will likely not yield such starkly divergent interpretations, but it nevertheless illustrates why accurately resolving the different noise contributions requires the time-dependent model developed here.

The misattribution of noise is particularly problematic in the development of kinetic models and analysis of experiments. Countless articles have presented stochastic simulations of noise in complex genetic circuits, and many appear to show strong quantitative agreement with experiments, but to our knowledge almost none have included duplication of the genes involved. One early study that did consider gene replication concluded that mRNA

relaxation contributed little to the overall noise, but this was based in part on an assumed mRNA half-life of 1 minute—considerably shorter than the mean for a bacteria [38] and in the regime where the corrections are predicted to be small. Returning to the hypothetical experiment described above, if a simple model of transcriptional regulation (such as Eq. 6.6) that did not account for gene duplication were used to fit the data, a modeler could arrive at estimates of k_{on} and k_{off}, for example, that deviate from the correct value by as much as 100%.

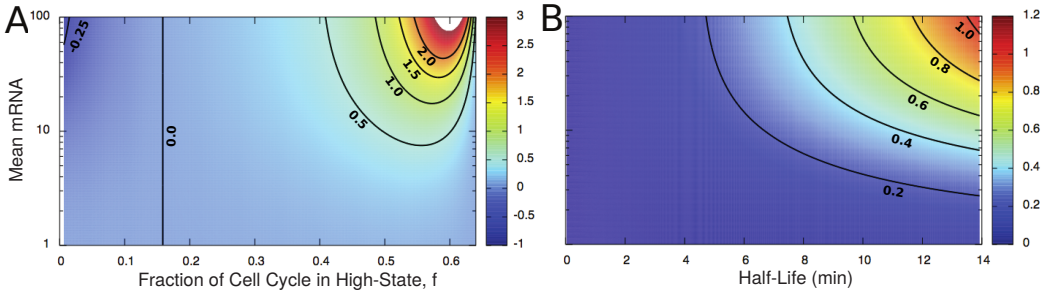


Figure 6.4: **Deviation of Time-Depedent and Time-Indepdendent Theories.** Error in the estimated Fano factor ($(F_{TI} - F_{TD})/F_{TI}$) when neglecting time-dependence of the mRNA relaxation as a function of: (A) the mean mRNA count and fraction of cell cycle after gene replication, and (B) mean mRNA and messanger half-life. Here a slow growing cell was considered (t_d ~70 minutes). In (A) the mRNA half-life was the average in *E coli* of 5.5 minutes. In (B) the fraction of the cell cycle after replication was taken to be 0.7. Scale bars indicate the value of the deviation. Contours are indicated with lines and the value along the contour denoted.

Because gene duplication-associated mRNA noise scales proportionally with the (mean) messenger expression level, the potential for its misattribution is greatest among highly expressed genes. In *E. coli*, these include a number of genes involved in key cellular processes like translation (including those encoding the ribosomal proteins), ATP synthesis (including

the ATP synthase genes), transcriptional regulation, and central metabolism (including the glycolytic genes *gapA* and *eno*) [53, 424]. The potential for noise misattribution is also related to f (the fraction of the cel cycle after gene duplication), and the messenger decay rate, k_d. Figure 6.4 shows the relative error between our time-dependent Fano factor expression and that of the time-independent theory (computed as $(F_{TI} - F_{TD})/F_{TI}$) for a cell doubling in 70 min. We see for highly expressed, long-lived transcripts the error can easily be $> 100\%$ while even in moderate cases the error can be in the range $20 - 50\%$ (most of this divergence comes from deviation of the TI model, as the TD model agrees well with simulation; see Fig. S6.21). Interestingly, because this error can change dramatically over a narrow range of values of f (*i.e.* $0.4 < f < 0.7$, see Fig. 6.4 A), and because f itself is a function of the cell's growth rate, small differences in cell doubling times can have a profound affect on the interpretation of mRNA noise. Taken altogether, these results indicate that the time-dependence of gene duplication and mRNA relaxation should not be ignored when either modeling stochastic gene expression or analyzing experimental data.

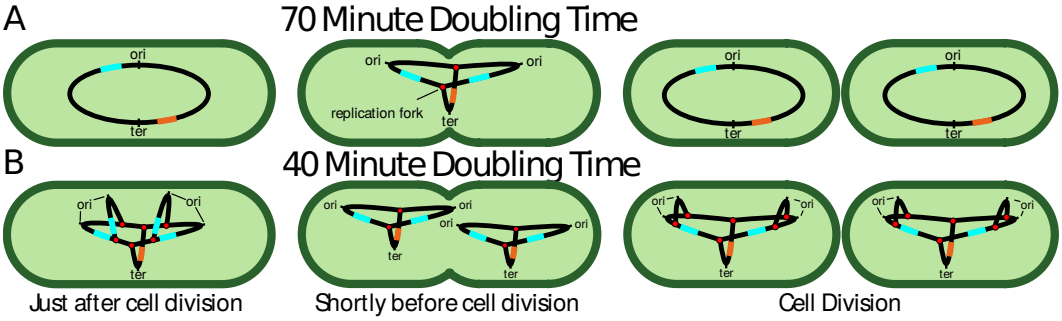


Figure 6.5: **Replication Schematics.** A schematic showing the replication of DNA containing one gene close to the origin (blue) and one close to the terminus (orange) at various timepoints in the cell cycle. Replication proceeds from the origin (*ori*) to the terminus (*ter*) and multiple replication forks (red dots) can exist simultaneously. Snapshots through the cell cycle from cells with doubling times (A) slower ($t_D = 70$ minutes) and (B) faster ($t_D = 40$ minutes) than the DNA replication time (45 minutes) are shown. For slow growing cells the initiation of replication occurs shortly after cell division and completes before the cell divides. For cells growing faster than the replication time, multiple copies of the genome must exist and therefore the number of replication forks can change dramatically throughout the cell cycle. The effect on gene count depends on the gene location; for instance a gene close to the origin is duplicated durign the same cell cycel that the replication is initiated, resulting in 2 or 4 copiles of the gene (A, middle). Conversely, a gene close to the terminus is replicated in the next cell cycle and only 1 or 2 copies can exist (B, right).

6.5 Supplementary Information

6.5.1 Derivation of the Fano Factor in the Case of Constitutive mRNA Expression

We consider the case where a single gene is present within a cell from time 0 to t_r—the gene replication time—and thereafter until the cell divides (at t_D) there are two copies. We are interested in computing how this gene-doubling

event impacts mRNA expression. We can begin by writing out the reaction network:

$$\begin{aligned} D_1 &\xrightarrow{k_{t,1}} D_1 + m \\ m &\xrightarrow{k_d} \emptyset \end{aligned} \tag{6.9}$$

for $t < t_r$, and:

$$\begin{aligned} D_1 &\xrightarrow{k_{t,1}} D_1 + m \\ D_2 &\xrightarrow{k_{t,2}} D_2 + m \\ m &\xrightarrow{k_d} \emptyset \end{aligned} \tag{6.10}$$

for $t > t_r$. These yield the chemical master equations (CMEs):

$$\begin{aligned} \partial_t P\left(m|t < t_r\right) &= k_{t,1} P\left(m-1|t\right) + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - k_{t,1} P\left(m|t\right) - k_d m P\left(m|t\right) \\ \partial_t P\left(m|t > t_r\right) &= \left(k_{t,1} + k_{t,2}\right) P\left(m-1|t\right) \\ &\quad + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - \left(k_{t,1} + k_{t,2}\right) P\left(m|t\right) - k_d m P\left(m|t\right) \end{aligned} \tag{6.11}$$

In the case where $k_{t,1} = k_{t,2} = k_t$, then the equation for $t > t_r$ above can be simplified as:

$$\begin{aligned} \partial_t P\left(m|t > t_r\right) &= 2k_t P\left(m-1|t\right) + k_d\left(m+1\right) P\left(m+1|t\right) \\ &\quad - 2k_t P\left(m|t\right) - k_d m P\left(m|t\right) \end{aligned} \tag{6.12}$$

To compute the time evolution of the mean $\bar{m}$ and variance of the mRNA distribution after the gene duplication event. We substituted the RHS of Eq. S6.12 into the definitions:

$$\begin{aligned}
\frac{d\bar{m}}{dt} &= \frac{d}{dt}\sum_{m=0}^{\infty} mP\left(m|t\right) \\
&= \sum_{m=0}^{\infty} m\frac{dP\left(m|t\right)}{dt} \\
&= \sum_{m=0}^{\infty} m\big[2k_tP\left(m-1|t\right) + k_d\left(m+1\right)P\left(m+1|t\right) \\
&\qquad - 2k_tP\left(m|t\right) - k_dmP\left(m|t\right)\big]
\end{aligned} \tag{6.13}$$

Evaluating each term individually:

$$\begin{aligned}
\sum_{m=0}^{\infty} m2k_tP\left(m-1|t\right) &= \sum_{y=-1}^{\infty}\left(y+1\right)2k_tP\left(y|t\right) \\
&= 0 + \sum_{y=0}^{\infty}\left(y+1\right)2k_tP\left(y|t\right) \\
&= 2k_t\left(1+\bar{y}\right) \\
&= 2k_t\left(1+\bar{m}\right)
\end{aligned} \tag{6.14}$$

$$\begin{aligned}\sum_{m=0}^{\infty} m k_d \left(m+1\right) P\left(m+1|t\right) &= \sum_{y=1}^{\infty} \left(y-1\right) y k_d P\left(y|t\right) \\ &= 0 + \sum_{y=0}^{\infty} \left(y-1\right) y k_d P\left(y|t\right) \\ &= k_d E\left[\left(y-1\right) y\right] \\ &= k_d E\left[\left(m-1\right) m\right] \\ &= k_d \left(E\left[m^2\right] - \bar{m}\right)\end{aligned} \tag{6.15}$$

$$\sum_{m=0}^{\infty} m 2 k_t P\left(m|t\right) = 2 k_t \bar{m} \tag{6.16}$$

and finally:

$$\sum_{m=0}^{\infty} m^2 k_d P\left(m|t\right) = k_d E\left[m^2\right] \tag{6.17}$$

Summing these with their appropriate signs yields:

$$\begin{aligned}\frac{d\bar{m}(t > t_r)}{dt} &= 2k_t + 2k_t\bar{m} + k_d E\left[m^2\right] - k_d\bar{m} - 2k_t\bar{m} - k_d E\left[m^2\right] \\ &= 2k_t - k_d\bar{m}\end{aligned} \tag{6.18}$$

while an identical argument gives:

$$\frac{d\bar{m}(t < t_r)}{dt} = k_t - k_d\bar{m} \tag{6.19}$$

The solution to these ODEs are straightforward. If gene duplication

occurs relatively early in the cell cycle, such that the mRNA count has time to equilibrate (to a value of $\frac{2k_t}{k_d}$) before the cell divides, then we can assume the mean mRNA count at the beginning of the cell cycle is approximately half this value, or $\frac{k_t}{k_d}$. Using this as an initial condition we can write down the mean as a function of time.

$$\bar{m}(t) = \begin{cases} \frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{6.20}$$

It should be noted that the assumption that the mean mRNA is approximately equilibrated at the end of the cell cycle is not strictly necessary (although it simplifies the resulting expressions enormously). In section S6.5.2, we give a more exact derivation that does not rely on this assumption.

We can now consider the evolution of the variance:

$$\begin{aligned} \frac{d\sigma_m^2}{dt} &= \frac{d}{dt}\left(E\left[m^2\right] - \bar{m}^2\right) \\ &= \frac{dE\left[m^2\right]}{dt} - 2\bar{m}\frac{d\bar{m}}{dt} \\ &= \frac{dE\left[m^2\right]}{dt} - 2\bar{m}\left(2k_t - k_d\bar{m}\right) \\ &= \left\{\sum_{m=0}^{\infty} m^2 \frac{dP\left(m|t\right)}{dt}\right\} - 2\bar{m}\left(2k_t - k_d\bar{m}\right) \\ &= \left\{\sum_{m=0}^{\infty} m^2 \big[2k_t P\left(m-1|t\right) + k_d\left(m+1\right)P\left(m+1|t\right) \right. \\ &\qquad \left. - 2k_t P\left(m|t\right) - k_d m P\left(m|t\right)\big]\right\} - 2\bar{m}\left(2k_t - k_d\bar{m}\right) \end{aligned} \tag{6.21}$$

As before, the terms in the summation are evaluated independently:

$$\begin{aligned}\sum_{m=0}^{\infty} m^2 2k_t P\left(m-1|t\right) &= \sum_{y=-1}^{\infty} \left(y+1\right)^2 2k_t P\left(y|t\right) \\ &= \sum_{y=0}^{\infty} \left(y+1\right)^2 2k_t P\left(y|t\right) \\ &= 2k_t \left(E\left[y^2\right] + 2\bar{y} + 1\right) \\ &= 2k_t \left(E\left[m^2\right] + 2\bar{m} + 1\right)\end{aligned} \tag{6.22}$$

$$\begin{aligned}\sum_{m=0}^{\infty} m^2 \left(m+1\right) k_d P\left(m+1|t\right) &= \sum_{y=1}^{\infty} y \left(y-1\right)^2 k_d P\left(y|t\right) \\ &= \sum_{y=0}^{\infty} y \left(y-1\right)^2 k_d P\left(y|t\right) \\ &= k_d \left(E\left[y^3\right] - 2E\left[y^2\right] + \bar{y}\right) \\ &= k_d \left(E\left[m^3\right] - 2E\left[m^2\right] + \bar{m}\right)\end{aligned} \tag{6.23}$$

$$\sum_{m=0}^{\infty} m^2 2k_t P\left(m|t\right) = 2k_t E\left[m^2\right] \tag{6.24}$$

and finally:

$$\sum_{m=0}^{\infty} m^3 k_d P\left(m|t\right) = k_d E\left[m^3\right] \tag{6.25}$$

Summing all these expressions together and simplifying gives:

$$\begin{aligned}\frac{d\sigma_m^2}{dt} &= 2k_t - 2k_d E\left[m^2\right] + k_d\bar{m} + 2k_d\bar{m}^2 \\ &= 2k_t - 2k_d \left(E\left[m^2\right] - \bar{m}^2\right) + k_d\bar{m} \\ &= 2k_t + k_d\bar{m} - 2k_d\sigma_m^2\end{aligned} \tag{6.26}$$

Substituting Eq. S6.20 and solving for $\sigma_m^2(t)$ when $t > t_r$ yields:

$$\sigma_m^2(t) = \frac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) + ce^{-2k_d t} \tag{6.27}$$

where c is an arbitrary integration constant. Noting that we are considering constitutively expressed mRNA for which we expect $m(t < t_r) \sim \text{Pois}(k_t/k_d)$, we can expect $\sigma_m^2(t_r) = k_t/k_d$. Using this in above as an initial condition yields:

$$\sigma_m^2(t) = \bar{m}(t) = \begin{cases} \frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_t}{k_d}\left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{6.28}$$

Interestingly, the mean and variance of the mRNA remain equal after the gene duplication event, indicating that the the mRNA remains Poisson–distributed. Although not necessary for the derivation at hand, substituting $P(m|t) = \text{Pois}(\bar{m}(t))$ into Eq. S6.11 shows that this is indeed the case.

Armed with these results, we can consider sampling the per-cell mRNA copy number of a population of cells. Assuming cells are sampled from across the cell cycle, we can write out the joint probability distribution for a randomly picked cell to have a given mRNA copy number:

$$P(m, t) = P(m|t)\, P(t) \tag{6.29}$$

where $P(m|t)$ is the distribution of m at a given time t along the cell cycle, and $P(t)$ represents the age distribution of cells in the population. For log-phase cells it has been shown that this distribution decays exponentially

with age [432, 434]. Ignoring cell-to-cell variability growth rate (which can be substantial [56]) and cell cycle duration, $P(t)$ can be given approximately by $\frac{2\,ln(2)}{t_D}\,2^{-t/t_D}$. From here, we can compute the probability that a cell will have m mRNA by simply marginalizing against the time variable:

$$P(m) = \int_0^{t_D} P(m,t)\,dt \tag{6.30}$$

Let's consider the expectation value of this distribution:

$$\begin{aligned} E[m] &= \sum_0^\infty mP(m) \\ &= \sum_0^\infty m \int_0^{t_D} P(m,t)dt \\ &= \sum_0^\infty m \int_0^{t_D} P(m|t)P(t)dt \\ &= \int_0^{t_D} \sum_0^\infty mP(m|t)P(t)dt \\ &= \int_0^{t_D} \bar{m}(t)\frac{2\,ln(2)}{t_D}\,2^{-t/t_D}dt \end{aligned} \tag{6.31}$$

Evaluating this yields:

$$E[m] = \frac{k_t}{k_d}\frac{1}{ln(2)+k_d t_D}[k_d t_D 2^{1-t_r/t_D} + ln(2)e^{-k_d(t_D-t_r)}] \tag{6.32}$$

Likewise, we can compute:

$$
\begin{aligned}
\text{Var}[m] &= \sum_{0}^{\infty} m^2 P(m) - E[m]^2 \\
&= \sum_{0}^{\infty} m^2 \int_0^{t_D} P(m|t)P(t)dt - E[m]^2 \\
&= \int_0^{t_D} \sum_{0}^{\infty} m^2 P(m|t)P(t)dt - E[m]^2 \\
&= \int_0^{t_D} (\sigma_m^2(t) + \bar{m}(t)^2)P(t)dt - E[m]^2 \\
&= \int_0^{t_D} \sigma_m^2(t)P(t)dt + \int_0^{t_D} \bar{m}(t)^2 P(t)dt - E[m]^2 \\
&= \int_0^{t_D} \bar{m}(t)P(t)dt + \int_0^{t_D} \bar{m}(t)^2 P(t)dt - E[m]^2 \\
&= E[m] + \int_0^{t_D} \bar{m}(t)^2 \frac{2\, ln(2)}{t_D}\, 2^{-t/t_D} dt - E[m]^2
\end{aligned}
\tag{6.33}
$$

We have already computed the functional form of $E[m]$, so all we have to do is evaluate the above integral:

$$
\begin{aligned}
\int_0^{t_D} \bar{m}(t)^2 \frac{2\, ln(2)}{t_D}\, 2^{-t/t_D} dt = (\frac{k_t}{k_d})^2 \Bigg[& (2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) \\
& + 4\, ln(2) \frac{e^{-k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} \\
& - ln(2) \frac{e^{-2k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D} \Bigg]
\end{aligned}
\tag{6.34}
$$

Packing this all up yields Var$[m]$:

$$\begin{aligned}\text{Var}[m] &= E[m] - E[m]^2 \\ &+ (\frac{k_t}{k_d})^2 \Bigg[(2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) \\ &+ 4\, ln(2) \frac{e^{-k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} - ln(2) \frac{e^{-2k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D} \Bigg]\end{aligned} \tag{6.35}$$

and hence:

$$\begin{aligned}\text{Fano}[m] &= \frac{\text{Var}[m]}{E[m]} \\ &= 1 - E[m] + \frac{1}{E[m]} (\frac{k_t}{k_d})^2 \Bigg[(2 - 2^{1-t_r/t_D}) - 4(1 - 2^{1-t_r/t_D}) \\ &+ 4\, ln(2) \frac{e^{-k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + k_d t_D} - ln(2) \frac{e^{-2k_d(t_D - t_r)} - 2^{1-t_r/t_D}}{ln(2) + 2k_d t_D} \Bigg]\end{aligned} \tag{6.36}$$

From here it is straightforward to cast these results in terms of the parameters $f = (t_D - t_r)/t_D$ (the fraction of cell cycle after the gene duplication event), and $\langle m \rangle_1 = k_t/k_d$ (the steady-state mean copy number prior to the gene duplication event):

$$
\begin{aligned}
E[m] &= \frac{\langle m \rangle_1}{1 + \frac{k_d t_D}{ln(2)}} \left[\frac{k_d t_D}{ln(2)} 2^f + e^{-k_d t_D f} \right] \\
\text{Var}[m] &= E[m] - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} \\
&\quad - ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} \Bigg] \\
\text{Fano}[m] &= 1 - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} \\
&\quad - ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} \Bigg]
\end{aligned}
\tag{6.37}
$$

If one were simply to assume a uniform distribution of ages, rather than the exponential distribution used above, the errors in $E[m]$ and Fano$[m]$ would be within approximately 6% for cells doubling in 40 min (and less than 8% for cells doubling in 70 min). The expressions that result, however, are considerably simpler (and thereby potentially more useful to a broad audience):

$$
\begin{aligned}
E[m] &= \langle m \rangle_1 \left[1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D} \right] \\
\text{Var}\,[m] &= E[m] - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2k_d t_D} \right] \\
\text{Fano}\,[m] &= 1 - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2f k_d t_D} - 7}{2k_d t_D} \right]
\end{aligned}
\tag{6.38}
$$

For this reason, we ultimately chose to include Eq. S6.38 in the main manuscript.

6.5.2 Relaxing the Assumption that the mRNA Counts Equilibrate Prior to Cell Division

Returning to Eqs. S6.18 & S6.19; in the prior section we solved for the mean copy number under the assumption that the mRNA counts have ample time to relax to the post-gene duplication steady state. For fast-growing cells with short doubling times, or when the gene of interest is duplicated near the end of a cell cycle, this assumption can prove untrue and can lead to disagreement between the analytical and simulated results (see for example Fig. S6.13). We can correct for this straightforwardly by introducing some initial mean mRNA count, $\bar{m}_0$, and solving for it by imposing the requirement that $\bar{m}_0 = \bar{m}(0) = \frac{1}{2}\bar{m}(t_D)$ (meaning the mean mRNA count after division is half what it was before division). We begin by writing down the mean before and after gene duplication:

$$\begin{aligned} \bar{m}(0 < t < t_r) &= (\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t} + \frac{k_t}{k_d} \\ \bar{m}(t_r < t < t_D) &= (\bar{m}(t_r) - 2\frac{k_t}{k_d})e^{-k_d(t-t_r)} + 2\frac{k_t}{k_d} \end{aligned} \tag{6.39}$$

Evaluating $\bar{m}(t_r)$ yields $(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} + \frac{k_t}{k_d}$ and, in turn, evaluating $\bar{m}(t_D)$ yields $[(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} - \frac{k_t}{k_d}]e^{-k_d(t_D - t_r)} + 2\frac{k_t}{k_d}$. Now simply imposing our boundary condition yields:

$$
\begin{aligned}
2\bar{m}_0 &= [(\bar{m}_0 - \frac{k_t}{k_d})e^{-k_d t_r} - \frac{k_t}{k_d}]e^{-k_d(t_D - t_r)} + 2\frac{k_t}{k_d} \\
\rightarrow \bar{m}_0 &= \frac{k_t}{k_d}[1 - \frac{e^{-k_d(t_D - t_r)}}{2 - e^{-k_d t_D}}]
\end{aligned}
\tag{6.40}
$$

This yields the exact solution for the mean:

$$
\bar{m}(t) = \begin{cases} \frac{k_t}{k_d}[1 - \frac{e^{-k_d(t_D - t_r)}}{2 - e^{-k_d t_D}} e^{-k_d t}] & 0 < t < t_r \\ \frac{k_t}{k_d}[2 - (1 + \frac{e^{-k_d t_D}}{2 - e^{-k_d t_D}})e^{-k_d(t - t_r)}] & t_r < t < t_D \end{cases}
\tag{6.41}
$$

Because, as noted in the previous section, Pois$(\bar{m}(t))$ solves the master equations for this problem, we can simply write $\sigma_m^2(t) = \bar{m}(t)$, although this could also be derived from Eq. S6.26 and similar arguments to those appearing above. Implicit in this, of course, is the assumption that the mRNA is Poisson–distributed after cell division; at least in the case of a perfectly unbiased division process this is easy to check. The probability that a daughter cell will contain m mRNAs immediately after division can be computed as:

$$
P_{\text{daughter}}(m) = \sum_{n=m}^{\infty} P(m|n) P_{\text{mother}}(n)
\tag{6.42}
$$

where $P_{\text{mother}}(n)$ represents the probability that the mother cell contains n mRNAs at division time, and $P(m|n)$ represents the probability that the daughter will contain m mRNA given that its mother contains n. If $P_{\text{mother}}(n) = \text{Pois}(\bar{n}(t_D))$ and cell division distributes mRNA with equal probabilities between the daughters we can write:

$$
\begin{aligned}
P_{\text{daughter}}(m) &= \sum_{n=m}^{\infty} \binom{n}{m} (\frac{1}{2})^n \frac{e^{-\bar{n}(t_D)} \bar{n}(t_D)^n}{n!} \\
&= \frac{e^{-\bar{n}(t_D)}}{m!} \sum_{n=m}^{\infty} \frac{(\frac{\bar{n}(t_D)}{2})^n}{(n-m)!} \\
&= \frac{e^{-\bar{n}(t_D)}}{m!} \sum_{k=0}^{\infty} \frac{(\frac{\bar{n}(t_D)}{2})^{k+m}}{k!} \\
&= \frac{e^{-\bar{n}(t_D)}}{m!} e^{\bar{n}(t_D)/2} (\frac{\bar{n}(t_D)}{2})^m \\
&= \text{Pois}(\frac{\bar{n}(t_D)}{2})
\end{aligned}
\tag{6.43}
$$

From this we see that the unbiased division of mRNA does indeed result in Poisson–distributed mRNA counts in the daughters.

Now we can compute the expectation value and variance for the messengers in a population of cells. Assuming cells are exponentially distributed yields:

$$
\begin{aligned}
E[m] &= \langle m \rangle_1 2^f \left[1 + \beta \frac{e^{-k_d t_D (1-f)} - 2^{1-f}}{1 + \frac{k_d t_D}{ln(2)}} + \gamma \frac{2^{-f} e^{-k_d t_D f} - 1}{1 + \frac{k_d t_D}{ln(2)}} \right] \\
\text{Var}[m] &= E[m] - E[m]^2 \\
&\quad + ln(2) \langle m \rangle_1^2 \left[2\beta^2 \frac{1 - 2^{f-1} e^{-2k_d t_D (1-f)}}{ln(2) + 2k_d t_D} - 4\beta \frac{1 - 2^{f-1} e^{-k_d t_D (1-f)}}{ln(2) + k_d t_D} \right. \\
&\quad + \frac{2}{ln(2)} (1 - 2^{f-1}) + \gamma^2 \frac{2^f - e^{-2k_d t_D f}}{ln(2) + 2k_d t_D} \\
&\quad \left. - 4\gamma \frac{2^f - e^{-k_d t_D f}}{ln(2) + k_d t_D} - \frac{4}{ln(2)} (1 - 2^f) \right] \\
\text{Fano}[m] &= 1 - E[m] \\
&\quad + ln(2) \frac{\langle m \rangle_1^2}{E[m]} \left[2\beta^2 \frac{1 - 2^{f-1} e^{-2k_d t_D (1-f)}}{ln(2) + 2k_d t_D} - 4\beta \frac{1 - 2^{f-1} e^{-k_d t_D (1-f)}}{ln(2) + k_d t_D} \right. \\
&\quad + \frac{2}{ln(2)} (1 - 2^{f-1}) + \gamma^2 \frac{2^f - e^{-2k_d t_D f}}{ln(2) + 2k_d t_D} \\
&\quad \left. - 4\gamma \frac{2^f - e^{-k_d t_D f}}{ln(2) + k_d t_D} - \frac{4}{ln(2)} (1 - 2^f) \right]
\end{aligned}
\tag{6.44}
$$

while assuming cells to be uniformly distributed yields:

$$
\begin{aligned}
E[m] &= \langle m \rangle_1 \left[1 + f + \frac{\beta}{k_d t_D}(e^{-k_d t_D (1-f)} - 1) + \frac{\gamma}{k_d t_D}(e^{-k_d t_D f} - 1) \right] \\
\text{Var}[m] &= E[m] - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \left[1 + 3f - \frac{4\beta}{2k_d t_D}(1 - e^{-k_d t_D (1-f)}) - \frac{8\gamma}{2k_d t_D}(1 - e^{-k_d t_D f}) \right. \\
&\quad \left. + \frac{\beta^2}{2k_d t_D}(1 - e^{-2k_d t_D (1-f)}) + \frac{\gamma^2}{2k_d t_D}(1 - e^{-2k_d t_D f}) \right] \\
\text{Fano}[m] &= 1 - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \left[1 + 3f - \frac{4\beta}{2k_d t_D}(1 - e^{-k_d t_D (1-f)}) - \frac{8\gamma}{2k_d t_D}(1 - e^{-k_d t_D f}) \right. \\
&\quad \left. + \frac{\beta^2}{2k_d t_D}(1 - e^{-2k_d t_D (1-f)}) + \frac{\gamma^2}{2k_d t_D}(1 - e^{-2k_d t_D f}) \right]
\end{aligned}
\tag{6.45}
$$

where:

$$
\begin{aligned}
\beta &= \frac{e^{-k_d t_D f}}{2 - e^{-k_d t_D}} \\
\gamma &= (1 + \frac{e^{-k_d t_D}}{2 - e^{-k_d t_D}})
\end{aligned}
\tag{6.46}
$$

These results show nearly exact agreement with simulation (see Figure S6.13). In the limit where f approaches 1 and t_D is large, these expressions reduce to those of Eqs. S6.37 & S6.38; the maximum difference ($\approx 17\%$) occurs when f is small, but we note that for values of f greater than 0.1 the difference between the above expressions and those of Eqs. S6.37 & S6.38 remains less than 8%. Because the expressions in equation S6.45 are considerably more complicated than those appearing in equation S6.38, and because they generally amount to a fairly small correction, we have not

included them in the main manuscript.

6.5.3 Corrections to the Fano Factor for the Case of Regulated mRNA Expression

When mRNA production is regulated (*e.g.* by a transcription factor) the dynamics of the system can be significantly more complicated. The behaviour of a single gene switching between "off" (dormant) and "on" (active, capable of producing mRNA) states has been studied on multiple occasions [41, 42, 416, 423, 426]. These analyses have shown that the mean and variance of the mRNA distribution at steady-state approach:

$$E_{ss}[m] = \frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \tag{6.47}$$

and:

$$\text{Var}_{ss}[m] = E_{ss}[m]\left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)}\right) \tag{6.48}$$

respectively.

As before, we are interested in computing the Fano factor in the case where a gene replication event occurs during the cell cycle. We saw previously that this quantity can be derived directly by integrating over time-dependent expressions for the mean and variance of the mRNA copy number. The mean is most easily studied in the continuum limit; we can write:

$$\frac{d\bar{m}(t)}{dt} = k_t\bar{g}(t) - k_d\bar{m}(t) \tag{6.49}$$

where $\bar{g}(t)$ represents the instantaneous mean number of "on" genes after the gene duplication event. As a first approximation, $\bar{g}$ might be assumed constant, and equal to $2k_{\text{on}}/(k_{\text{on}} + k_{\text{off}})$. This would be reasonable in the limit where the gene switching is fast such that $\bar{g}$ relaxes to its new steady state quickly compared to the time required for $\bar{m}(t)$ to relax to its new steady state. This assumption immediately yields:

$$\bar{m}(t) = \begin{cases} \frac{k_t}{k_d} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{k_t}{k_d} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \left(2 - e^{k_d(t_r - t)}\right) & t_r < t < t_D \end{cases} \tag{6.50}$$

Not surprisingly, the mean expression is identical to the unregulated case, except that it is scaled by $\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$.

We can now turn our attention to the estimating $\sigma_m^2(t)$. This is non-trivial, and we will not attempt a complete derivation here. But Eq. S6.48 indicates that the variance before and long after the duplication event should be proportional to the mean; if, as a first approximation, we were to simply assume that the dynamics of the variance occur on a similar time-scale as the dynamics of the mean, then we could immediately write:

$$\sigma_m^2(t) \approx \bar{m}(t) \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)}\right) \tag{6.51}$$

which yields:

$$\text{Var}[m] = E[m] \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)}\right) - E[m]^2 + \int_0^{t_D} \bar{m}(t)^2 P(t) dt \tag{6.52}$$

Evaluating this assuming an exponential age distribution yields:

$$
\begin{aligned}
E[m] &= \frac{\langle m \rangle_1}{1 + \frac{k_d t_D}{ln(2)}} \left[\frac{k_d t_D}{ln(2)} 2^f + e^{-k_d t_D f} \right] \\
\text{Var}[m] &= E[m] \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} \\
&\quad - ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} \Bigg] \\
\text{Fano}[m] &= \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \Bigg[(2 - 2^f) - 4(1 - 2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} \\
&\quad - ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} \Bigg]
\end{aligned}
\tag{6.53}
$$

where

$$
\langle m \rangle_1 = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \frac{k_t}{k_d} \tag{6.54}
$$

Evaluating assuming a uniform distribution yields:

$$
\begin{aligned}
E[m] &= \langle m \rangle_1 \left[1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D} \right] \\
\text{Var}\,[m] &= E[m] \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m]^2 \\
&\quad + \langle m \rangle_1^2 \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2 f k_d t_D} - 7}{2 k_d t_D} \right] \\
\text{Fano}\,[m] &= \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)} \right) - E[m] \\
&\quad + \frac{\langle m \rangle_1^2}{E[m]} \left[1 + 3f + \frac{8e^{-f k_d t_D} - e^{-2 f k_d t_D} - 7}{2 k_d t_D} \right]
\end{aligned}
\tag{6.55}
$$

Comparison with Eqs. S6.37 & S6.38 shows that these are functionally very similar to the unregulated case but with an additional term, $\frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}$, and, as noted above, the replacement $\langle m \rangle_1 \rightarrow \frac{k_t}{k_d} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$.

6.5.4 Further Corrections for Cases in which $k_{\text{on}}, k_{\text{off}} \lesssim k_d$

When the gene state switching rates, k_{on} and k_{off}, are not faster than k_d, we might expect that some additional refinements are in order. The first of which would be that the dynamics of $\bar{g}(t)$ appearing in Eq. S6.49 ought not be ignored.

We can write down a differential equation for the $\bar{g}(t)$ after the gene duplication:

$$
\frac{d\bar{g}(t)}{dt} = (2 - \bar{g}(t)) k_{\text{on}} - \bar{g}(t) k_{\text{off}} \tag{6.56}
$$

for which the solution is:

$$\bar{g}(t) = \frac{2k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} + ce^{-(k_{\text{on}}+k_{\text{off}})t} \tag{6.57}$$

where c is an arbitrary integration constant. There are a few things to consider before we decide on an initial condition. Genes are replicated by a large protein complex that sweeps along the DNA, unzipping it and replicating both strands as it goes. Any transcription factors bound to the original gene copy's promoter region would have been unbound by the replication complex, and so both genes start off in an unbound state at time t_r. This can mean one of two things—if the transcription factor was a repressor, then both genes would begin "on" ($\bar{g}(t_r) = 2$), while if it were an activator, both genes would begin "off" ($\bar{g}(t_r) = 0$). This yields:

$$\bar{g}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{2k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \left(1 - e^{(k_{\text{on}}+k_{\text{off}})(t_r-t)}\right) & t_r < t < t_D \end{cases} \tag{6.58}$$

if the regulator is an activator, and:

$$\bar{g}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} & 0 < t < t_r \\ \frac{2}{k_{\text{on}}+k_{\text{off}}} \left(k_{\text{on}} + k_{\text{off}} e^{(k_{\text{on}}+k_{\text{off}})(t_r-t)}\right) & t_r < t < t_D \end{cases} \tag{6.59}$$

if the regulator is a repressor.

We can insert these into Eq. S6.49 and solve it yielding:

$$\bar{m}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d}\Bigg[\left(2 - e^{k_d(t_r-t)}\right) & \\ + \frac{2k_d}{k_{\text{on}}+k_{\text{off}}-k_d}\left(e^{(k_{\text{on}}+k_{\text{off}})(t_r-t)} - e^{k_d(t_r-t)}\right)\Bigg] & t_r < t < t_D \end{cases} \tag{6.60}$$

if the regulator is an activator, and:

$$\bar{m}(t) = \begin{cases} \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d} & 0 < t < t_r \\ \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\frac{k_t}{k_d}\Bigg[\left(2 - e^{k_d(t_r-t)}\right) & \\ - \frac{2k_d}{k_{\text{on}}+k_{\text{off}}-k_d}\frac{k_{\text{off}}}{k_{\text{on}}}\left(e^{(k_{\text{on}}+k_{\text{off}})(t_r-t)} - e^{k_d(t_r-t)}\right)\Bigg] & t_r < t < t_D \end{cases} \tag{6.61}$$

if the regulator is a repressor.

From here, evaluating $E[m]$ and Var$[m]$ is straightforward, if somewhat laborious. In the end, the functional forms they take are not particularly illuminating, but we have included them here for the sake of completeness:

$$
\begin{aligned}
E_{\text{act}}[m] = & \frac{1}{t_D}\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \\
& \times \left(t_r + \frac{2k_d}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}-k_d)} \left(1 - e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right) \right. \\
& \left. + \frac{k_{\text{on}}+k_{\text{off}}+k_d}{k_{\text{on}}+k_{\text{off}}-k_d}\frac{1}{k_d}\left(e^{-k_d(t_D-t_r)}-1\right) + 2(t_D-t_r) \right)
\end{aligned}
\tag{6.62}
$$

and:

$$
\begin{aligned}
\text{Var}_{\text{act}}[m] = & \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) E_{\text{act}}[m] - E_{\text{act}}[m]^2 \\
& + eta\frac{t_r}{t_D}\left(\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\right)^2 + \frac{k_t^2 k_{\text{on}}^2}{2t_D k_d^2 (k_{\text{on}}+k_{\text{off}})^2 (k_{\text{on}}+k_{\text{off}}-k_d)^2} \\
& \times \left(\frac{4k_d^2}{(k_{\text{on}}+k_{\text{off}})}\left(1 - e^{-2(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right) \right. \\
& + \frac{(k_{\text{on}}+k_{\text{off}}+k_d)^2}{k_d}\left(1 - e^{-2k_d(t_D-t_r)}\right) \\
& + \frac{k_d^2 - (k_{\text{on}}+k_{\text{off}})^2}{k_d}\left(1 - e^{-k_d(t_D-t_r)}\right) \\
& - \frac{8k_d(2k_d - k_{\text{on}} - k_{\text{off}})}{k_{\text{on}}+k_{\text{off}}} \\
& + 8k_d\left(e^{-(k_{\text{on}}+k_{\text{off}}+k_d)(t_D-t_r)} + 2\frac{k_d - k_{\text{on}} - k_{\text{off}}}{k_{\text{on}}+k_{\text{off}}} e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)} \right) \\
& \left. + 8(k_{\text{on}}+k_{\text{off}}-k_d)^2(t_D-t_r) \right)
\end{aligned}
\tag{6.63}
$$

for activation-type regulation, and:

$$
\begin{aligned}
E_{\text{rep}}[m] = & \frac{1}{t_D}\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} \\
& \times \left(t_r + \frac{2k_d k_{\text{off}}\left(e^{-(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}-1\right)}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}-k_d)k_{\text{on}}} \right. \\
& \left. + \frac{k_d(2k_{\text{off}}+k_{\text{on}}) - k_{\text{on}}(k_{\text{off}}+k_{\text{on}})}{k_d k_{\text{on}}(k_{\text{off}}+k_{\text{off}}-k_d)}\left(1-e^{-k_d(t_D-t_r)}\right) + 2(t_D-t_r) \right)
\end{aligned}
\tag{6.64}
$$

and:

$$
\begin{aligned}
\text{Var}_{\text{rep}}[m] = & \left(1 + \frac{k_t k_{\text{off}}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\right) E_{\text{act}}[m] - E_{\text{act}}[m]^2 \\
& + \frac{t_r}{t_D}\left(\frac{k_t}{k_d}\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}\right)^2 + \frac{k_t^2}{2t_D k_d^2 (k_{\text{on}}+k_{\text{off}})^2 (k_{\text{on}}+k_{\text{off}}-k_d)^2} \\
& \times \left(\frac{4k_d^2 k_{\text{off}}^2}{(k_{\text{on}}+k_{\text{off}})}\left(1-e^{-2(k_{\text{on}}+k_{\text{off}})(t_D-t_r)}\right) \right. \\
& + \frac{(k_{\text{on}}(k_{\text{on}}+k_{\text{off}}) - k_d(2k_{\text{off}}+k_{\text{on}}))^2}{k_d}\left(1-e^{-2k_d(t_D-t_r)}\right) \\
& + \frac{8k_{\text{on}}\left(k_d^2\left(2k_{\text{off}}+k_{\text{on}}\right) - 2k_d\left(k_{\text{off}}+k_{\text{on}}\right)^2 + k_{\text{on}}\left(k_{\text{off}}+k_{\text{on}}\right)^2\right)}{k_d} \\
& \times \left(e^{-k_d(t_D-t_r)}-1\right) \\
& + \frac{8k_d k_{\text{off}}\left(-k_d\left(3k_{\text{off}}k_{\text{on}} + 2k_{\text{off}}^2 + k_{\text{on}}^2\right) + 2k_d^2 k_{\text{on}} - k_{\text{on}}\left(k_{\text{off}}+k_{\text{on}}\right)^2\right)}{\left(k_{\text{off}}+k_{\text{on}}\right)\left(k_d+k_{\text{off}}+k_{\text{on}}\right)} \\
& - \frac{8k_d k_{\text{off}} e^{-(k_{\text{on}}+k_{\text{off}}+k_d)(t_D-t_r)}}{(k_{\text{on}}+k_{\text{off}})(k_{\text{on}}+k_{\text{off}}+k_d)}\left[2k_d^2 k_{\text{on}} e^{k_d(t_D-t_r)} \right. \\
& \left. - k_{\text{on}}(k_{\text{on}}+k_{\text{off}})^2\left(2e^{k_d(t_D-t_r)}-1\right) - k_d\left(2k_{\text{off}}^2 + 3k_{\text{off}}k_{\text{on}} + k_{\text{on}}^2\right)\right] \\
& \left. + 8k_{\text{on}}^2(k_{\text{on}}+k_{\text{off}}-k_d)^2(t_D-t_r)\right)
\end{aligned}
\tag{6.65}
$$

for repression-type regulation.

6.5.5 Corrections to the Fano Factor Arising from Variability in RNAP Copy Number

We consider how cell-to-cell variability in RNA polymerase (RNAP) copy numbers can impact the Fano factor of a gene that doubles during the cell cycle. In our analysis thus far, we have considered a single constant transcription rate, k_t. If we assume that this rate is proportional to the number of RNAPs available to transcribe a gene, then we can simply promote k_t to a random variable and analyse its effect on our earlier results. For simplicity, we return to the case of constitutive expression, and specifically to our considerations of $P(m)$ (see Eqs. S6.29 and S6.30). In that case we had assumed k_t was fixed and derived the mean and variance of $P(m|t)$ at every t; now we assume k_t is random and realize our expressions actually give the mean and variance of $P(m|t, k_t)$. From here we simply write:

$$\begin{aligned} P(m) &= \int_0^\infty \int_0^{t_D} P(m, t, k_t)\, dt\, dk_t \\ &= \int_0^\infty \int_0^{t_D} P(m|t, k_t) P(t) P(k_t)\, dt\, dk_t \end{aligned} \tag{6.66}$$

Now, evaluating $E[m]$ follows the same logic as before:

$$
\begin{aligned}
E[m] = \sum_0^\infty mP(m) &= \sum_0^\infty m \int_0^\infty dk_t \int_0^{t_D} dt P(m|t,k_t)P(t)P(k_t) \\
&= \int_0^\infty \int_0^{t_D} \sum_0^\infty mP(m|t,k_t)P(t)P(k_t)\, dt\, dk_t \\
&= \int_0^\infty \int_0^{t_D} \bar{m}(t,k_t)P(t)P(k_t)\, dt\, dk_t \\
&= \int_0^\infty E[m|k_t]P(k_t)\, dk_t \\
&= E_{k_t}[E[m|k_t]] \\
&\approx E[m|\bar{k_t}] + \frac{1}{2}\left(\frac{\partial^2 E[m|k_t]}{\partial k_t^2}\right)\Bigg|_{\bar{k_t}} \text{Var}[k_t] \\
&= E[m|\bar{k_t}]
\end{aligned}
\tag{6.67}
$$

where $E[m|k_t]$ represents $E[m]$ (given by Eq. S6.32) evaluated at a specific value of k_t, $E_{k_t}[f(k_t)]$ represents the expectation value of $f(k_t)$ over k_t, and $\bar{k_t}$ represents the mean value of k_t. Note that the second to last line follows from a Taylor expansion of $E[m|k_t]$ about $E[m|\bar{k_t}]$.

Likewise we can do the same type of analysis for Var$[m]$:

$$
\begin{aligned}
\text{Var}[m] &= \sum_0^\infty m^2 P(m) - E[m]^2 \\
&= \sum_0^\infty m^2 \int_0^\infty dk_t \int_0^{t_D} dt P(m|t,k_t) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \int_0^{t_D} \sum_0^\infty m^2 P(m|t,k_t) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \int_0^{t_D} \left(\sigma_m^2(t,k_t) + \bar{m}^2(t,k_t)\right) P(t) P(k_t)\, dt\, dk_t - E[m]^2 \\
&= \int_0^\infty \left(\text{Var}[m|k_t] + E[m|k_t]^2\right) P(k_t)\, dk_t - E[m]^2 \\
&= E_{k_t}\left[\text{Var}[m|k_t] + E[m|k_t]^2\right] - E[m]^2 \\
&= E_{k_t}\left[\text{Var}[m|k_t]\right] + E_{k_t}[E[m|k_t]^2] - E[m]^2 \\
&\approx \text{Var}[m|\bar{k_t}] + \frac{1}{2}\left(\frac{\partial^2 \text{Var}[m|k_t]}{\partial k_t^2}\right)\Bigg|_{\bar{k_t}} \text{Var}[k_t] \\
&+ E[m|\bar{k_t}]^2 + \frac{1}{2}\left(\frac{\partial^2 E[m|k_t]^2}{\partial k_t^2}\right)\Bigg|_{\bar{k_t}} \text{Var}[k_t] - E[m]^2
\end{aligned}
\tag{6.68}
$$

which ultimately yields:

$$
\text{Var}[m] \approx \text{Var}[m|\bar{k_t}] + \frac{\nu \text{Var}[k_t]}{k_d^2} \tag{6.69}
$$

where

$$\nu = \begin{cases} (2-2^f) - 4(1-2^f) + 4\, ln(2)\frac{e^{-k_d t_D f} - 2^f}{ln(2)+k_d t_D} & \\ -ln(2)\frac{e^{-2k_d t_D f} - 2^f}{ln(2)+2k_d t_D} & \text{when } P(t) \text{ is exponential} \\ 1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases} \tag{6.70}$$

and therefore the Fano factor simply picks up an additive correction:

$$\text{Fano}[m] \approx \text{Fano}[m|\bar{k_t}] + \frac{\nu \text{Var}[k_t]}{E[m]k_d^2} \tag{6.71}$$

for the case of constitutive mRNA expression.

It is worth noting that nothing about this analysis is specific to k_t; the same types of arguments can be made in order to account for variability in any parameters, including the mRNA degradation rate (k_d), the cell cycle length (t_D), or the gene replication time (t_r). In the next section we will use a similar argument to consider how variability in transcription factor copy number affects regulated mRNA expression.

6.5.6 Corrections to the Fano Factor Arising from Variability in Transcription Factor Copy Number

We return to our earlier consideration of regulated mRNA expression (see Section S6.5.3). We had previously considered a model wherein the gene can be in either an "on" state capable of being transcribed or an "off" state which is incapable of being transcribed. We might assume that the transition

between states is mediated by the binding of a transcription factor (TF). As we did previously for k_t, we can promote k_{on} (if the TF is an activator) or k_{off} (if the TF is a repressor) to a random variable which can be assumed to vary from cell-to-cell. We can then compute how this effects the mean mRNA copy number. If, for example, we assume the TF is an activator we can write:

$$\begin{aligned} E[m] &\approx E[m|\bar{k_{\text{on}}}] + \frac{1}{2}\left(\frac{\partial^2 E[m|k_{\text{on}}]}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] \\ &= E[m|\bar{k_{\text{on}}}]\left[1 - \frac{k_{\text{off}}\text{Var}[k_{\text{on}}]}{\bar{k_{\text{on}}}(\bar{k_{\text{on}}} + k_{\text{off}})^2}\right] \end{aligned} \tag{6.72}$$

Evaluating the variance is tedious, and likely best performed using a computer algebra system:

$$
\begin{aligned}
\text{Var}[m] \approx\ & \text{Var}[m|\bar{k_{\text{on}}}] + \frac{1}{2}\left(\frac{\partial^2 \text{Var}[m|k_{\text{on}}]}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] \\
& + E[m|\bar{k_{\text{on}}}]^2 + \frac{1}{2}\left(\frac{\partial^2 E[m|k_{\text{on}}]^2}{\partial k_{\text{on}}^2}\right)\Bigg|_{\bar{k_{\text{on}}}} \text{Var}[k_{\text{on}}] - E[m]^2 \\
& = \text{Var}[m|\bar{k_{\text{on}}}] + E[m|\bar{k_{\text{on}}}]^2\left[1 - \left(1 - \frac{k_{\text{off}}\text{Var}[k_{\text{on}}]}{\bar{k_{\text{on}}}(\bar{k_{\text{on}}} + k_{\text{off}})^2}\right)^2\right] \\
& - \frac{\text{Var}[k_{\text{on}}]k_{\text{off}}k_t}{k_d^2(\bar{k_{\text{on}}} + k_{\text{off}})^4(k_d + \bar{k_{\text{on}}} + k_{\text{off}})^3} \\
& \times \Bigg[2\bar{k_{\text{on}}}^4\nu k_t - k_{\text{off}}^4\nu k_t + 3k_d^2\eta k_{\text{off}}^3 \\
& + 3k_d^3\eta k_{\text{off}}^2 + 3k_d^2\eta\bar{k_{\text{on}}}^3 + 3k_d^3\eta\bar{k_{\text{on}}}^2 \\
& + k_d\eta k_{\text{off}}^4 + k_d^4\eta k_{\text{off}} + k_d\eta\bar{k_{\text{on}}}^4 \\
& + k_d^4\eta\bar{k_{\text{on}}} + 4k_d\eta k_{\text{off}}\bar{k_{\text{on}}}^3 + 4k_d\eta k_{\text{off}}^3\bar{k_{\text{on}}} \\
& + 6k_d^3\eta k_{\text{off}}\bar{k_{\text{on}}} + 3k_d\eta k_{\text{off}}^3 k_t + 2k_d^3\eta k_{\text{off}}k_t \\
& - 3k_d\eta\bar{k_{\text{on}}}^3 k_t - k_d^3\eta\bar{k_{\text{on}}}k_t - 3k_d k_{\text{off}}^3\nu k_t \\
& - k_d^3 k_{\text{off}}\nu k_t + 6k_d\bar{k_{\text{on}}}^3\nu k_t + 2k_d^3\bar{k_{\text{on}}}\nu k_t \\
& + 5k_{\text{off}}\bar{k_{\text{on}}}^3\nu k_t - k_{\text{off}}^3\bar{k_{\text{on}}}\nu k_t + 6k_d\eta k_{\text{off}}^2\bar{k_{\text{on}}}^2 \\
& + 9k_d^2\eta k_{\text{off}}\bar{k_{\text{on}}}^2 + 9k_d^2\eta k_{\text{off}}^2\bar{k_{\text{on}}} + 5k_d^2\eta k_{\text{off}}^2 k_t \\
& - 3k_d^2\eta\bar{k_{\text{on}}}^2 k_t - 3k_d^2 k_{\text{off}}^2\nu k_t + 6k_d^2\bar{k_{\text{on}}}^2\nu k_t \\
& + 3k_{\text{off}}^2\bar{k_{\text{on}}}^2\nu k_t - 3k_d\eta k_{\text{off}}\bar{k_{\text{on}}}^2 k_t + 3k_d\eta k_{\text{off}}^2\bar{k_{\text{on}}}k_t \\
& + 2k_d^2\eta k_{\text{off}}\bar{k_{\text{on}}}k_t + 9k_d k_{\text{off}}\bar{k_{\text{on}}}^2\nu k_t + 3k_d^2 k_{\text{off}}\bar{k_{\text{on}}}\nu k_t\Bigg]
\end{aligned}
\tag{6.73}
$$

where:

$$\nu = \begin{cases} (2-2^f) - 4(1-2^f) + 4\, ln(2) \frac{e^{-k_d t_D f} - 2^f}{ln(2) + k_d t_D} & \\ -ln(2) \frac{e^{-2k_d t_D f} - 2^f}{ln(2) + 2k_d t_D} & \text{when } P(t) \text{ is exponential} \\ 1 + 3f + \frac{8e^{-fk_d t_D} - e^{-2fk_d t_D} - 7}{2k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases} \tag{6.74}$$

and:

$$\eta = \begin{cases} \frac{1}{1 + \frac{k_d t_D}{ln(2)}} \left[\frac{k_d t_D}{ln(2)} 2^f + e^{-k_d t_D f} \right] & \text{when } P(t) \text{ is exponential} \\ 1 + f + \frac{e^{-fk_d t_D} - 1}{k_d t_D} & \text{when } P(t) \text{ is uniform} \end{cases} \tag{6.75}$$

This is, obviously, a somewhat long and cumbersome expression, but we can nonetheless estimate the size of the corrections. If we assume $k_d = 0.126$ min^{-1}, $k_t = 10k_d$, and that there are on average 10 copies of the TF per cell (which, assuming the TF is in the extrinsic noise limit where Var[TF]$/E$[TF]$^2 \approx 0.1$ [424], yields an estimate of Var$[k_{\text{on}}] \approx 0.1 \times \bar{k_{\text{on}}}^2$), we can compute the corrections to the mean and Fano factor spanning a range of values of $\bar{k_{\text{on}}}$ and k_{off} (from 10^{-3} to 10^2 min^{-1} in both rates). Accounting for TF variability generally resulted in a small decrease in the value of $E[m]$. Over the range of kinetic parameters studied, the largest change to the mean mRNA copy number computed was only around 3%. TF variability made a marginally larger impact on the Fano factor values, but these changes were highly dependent on the values of $\bar{k_{\text{on}}}$ and k_{off}. We found that when $\bar{k_{\text{on}}} << k_{\text{off}}$, the Fano factor increased by approximately $E[m]/10$—a contri-

bution of similar magnitude to that stemming from RNAP variability—but when $k_{\text{on}}^{-} \sim k_{\text{off}}$ we find that this contribution drops to below 3% of $E[m]$. When $k_{\text{on}}^{-} >> k_{\text{off}}$ the correction becomes vanishingly small. Importantly, these results indicate that TF variability generally imparts less mRNA noise than does RNAP variability.

6.5.7 Comparison Between Different Models Considering Gene Copy Number Variation

In order to show that mRNA relaxation dynamics play an important role and that gene replication should be handle explicitly, we compared to previous treatments of gene copy number effect. Two studies have examined the contributions to transcriptional noise arising from of variations in gene copy number [49, 425]. In one model, the "constant DNA model", the average number of gene copies across a population of cells is computed and each simulated cell is assumed to have this copy number over all time [425]. In this case, the mean mRNA copy number scales linearly with gene copy number and for constitutively expressed genes the Fano factor remains unitary. This is due to the fact that the addition of multiple Poisson variables (*i.e.* the mRNA produced from each gene), yields a Poisson variable. Therefore, simulating a population with an average gene copy of 1 or 3 does not change the observed noise (see Fig. S6.6, blue bars).

In the second model, the "weighted DNA model" [49], each replicating cell's gene copy number is constant over the cell cycle; however, the copy number for each cell is drawn from the distribution of copy numbers

observed in a population of cells. In this case, the theory of Cooper and Helmstetter [430] can be used to calculate the probability of having a particular gene copy number from the doubling time and gene location by taking the fraction of the cell cycle in which that number count exists (see Tables in Figs. S6.9C and S6.8C). We simulated cells in slow growth (70 min doubling time) and a fast growth (40 min doubling time) conditions, with genes located 10% and 90% from replication origin. Consistent with reported by Jones *et al.*, under all four conditions, Fano factors are increased to different extend, demonstrating approaches using the constant DNA model are qualitatively wrong (Fig. S6.6, yellow bars). When we compared the results obtained when simulating DNA replication explicitly, it became apparent that the weighted model overestimates the noise from gene replication (compare red to yellow bars in Fig. S6.6B). Including explicit replication and mRNA relaxation dynamics in fact results in noise that is consistently lower than the weighted model by a significant amount (Fig. S6.6, red bars). Simulated mRNA distributions for a wide range of mRNA locations at two different doubling times demonstrate that the weighted DNA model is consistently quantitatively, and even qualitatively incapable of capturing the observed noise from these more realistic simulations (see, for example, Fig. 2 in the main manuscript, Figs. S6.9, S6.8, and S6.11 blue lines). On the other hand, the mean and Fano factor as well as mRNA distributions computed via the time-dependent theory developed in this paper are both qualitatively and quantitatively more correct when compared to exact simulations in almost every scenario studied, demonstrating that mRNA relaxation dynamics,

which constitute a significant portion of the overall cell cycle, impact the statistics of observable mRNA copy numbers. As a concrete example, to reach the new steady–state mRNA level after DNA replication (defined by relaxation to within 1σ of mean), it takes $\sim$6.4 min, $\sim$16% of a 40-min cell cycle.

6.5.8 Simulated and Analytical Distributions for Constitutively Expressed Genes

Expanded versions of Fig. 2 are shown in Figs. S6.8 and S6.9 that include values the average gene count as a function of the distance from *ori* to *ter*. In addition to the distributions shown in these figures, we have computed the exact distributions for constitutively expressed genes located at positions spaced every 5% of the way from origin to terminus for fast- and slow-growing cells. Distributions were computed by integrating Eq. S6.30 assuming a Poisson-distributed mRNA number at each point along the cell cycle, where the mean and variance of the distribution is taken to be Eq. S6.20. Resulting distributions are shown in Figs. S6.10 and S6.11.

The TD theory assuming the mRNA relaxes to steady–state before cell division is not exact for all values of f as demonstrated by the relatively poor agreement for the 40 minute doubling time case for genes that duplicate near cell division (see, for instance, Fig. S6.12 and Fig. S6.10). When a gene duplicates very close to cell division, the mRNA has insufficient time to relax to the high steady-state, and therefore after cell division, the average level does not represent the low steady-state. In fact, the gene must relax

after cell division up to the low steady-state, prior to gene duplication. This phenomenon can be seen for a gene located 60% of the way from origin to terminus in Fig. S6.13 wherein there are two clear relaxations. Our model does not capture this behaviour, however, we derived slightly more involved equations that take this into account (Eqs. S6.44-6.46).

$$D_{KL}\left(P||Q\right) = \sum_{m} P\left(m\right) \ln\left(\frac{P\left(m\right)}{Q\left(m\right)}\right) \tag{6.76}$$

6.5.9 Numerical vs. Experimental Distributions

To assess the quality of the theory on real world data, it was applied to the various mutation studies reported in Jones et al. and compared against their experimental data [49]. Their data is associated with a gene that spends 1/3 of the cell cycle before gene duplication (f=0.67). The mean of the experimental data was taken to be the time-averaged mean $\langle m \rangle$ of Eq. S6.37, which was used to compute the mRNA for the low state, $\langle m \rangle_1$ as:

$$\langle m \rangle_1 = \frac{\langle m \rangle}{1 + f + \frac{e^{-f k_d t_D} - 1}{k_d t_D}} \tag{6.77}$$

The mRNA half-live and the cell doubling time must also be defined in order to compute the theoretical distributions. We took the mRNA half-life to be 5.5 minutes as done in the main text. Jones et al. grew their *E. coli* in M9 minimal salts media supplemented with 0.5% glucose so t_D was taken as 40 minutes (Reshes *et al.* reported an *E. coli* doubling time of 38$\pm$1 min for cells grown in M9 salts+0.4% glucose [438]). Using these parameters for

k_d and t_D, $\langle m \rangle_1$ was computed via Eq. S6.77, and the exact distributions of mRNA from the time-dependent theory were computed by integrating Eq. S6.30 considering only the case of constitutively expressed mRNA.

In the case of the time-independent theory, Eq. S6.77 reduces to:

$$\langle m \rangle_1 = \frac{\langle m \rangle}{1+f} \tag{6.78}$$

as found in Jones et al. [49]. We used this mean to compute the time–independent distribution as:

$$P\left(m\right) = f\text{Pois}\left(2\langle m \rangle_1\right) + (1-f)\text{Pois}\left(\langle m \rangle_1\right) \tag{6.79}$$

Figure S6.14 shows the comparison of these distributions to experiments. The resulting time-dependent distributions better represent the data than do those of the time-independent theory; capturing the shape both qualitatively and quantitatively. As discussed in the main text (see Figure 4), the time-dependent theory becomes more important as the mean mRNA becomes large ($\langle m \rangle_1 < 1.0$), and this holds true when comparing to experimental data. However, the time-dependent theory is even quantitatively better for experiments with mean mRNA smaller than 1 (see Figure S6.14).

In order to quantify the agreement, we compute the mean-squared deviation (MSD) between the computed and experimental distributions as:

$$MSD = \frac{1}{N}\sum_{m=0}^{N}\left(P_{\text{theory}}\left(m\right) - P_{\text{exp}}\left(m\right)\right)^2 \tag{6.80}$$

The results are shown in Fig. S6.15A. Indeed the MSD was smaller using

the time-dependent theory in every case, sometimes up to a factor of 10x, verifying that the theory is appropriate. A KL-divergence, computed neglecting contributions where the experimental probability is zero, qualitatively shows the same picture (Fig. S6.15B).

As an independent validation, we compared the experimental distribution, computed by pooling the data from two independent experiments, of the messenger RNA *ptsG* acquired via smFISH of *E. coli* cells growing in 0.2% glucose (see [156] for experimental methods). In this case, the gene is located 25% of the way form the origin to the terminus and the cells had doubling times $\sim$40 minutes; therefore f=0.34375. The degradation rate k_d was previously measured to be 0.246$\pm$0.049 min^{-1} [156]. This time, the theory based on constitutive expression cannot capture the mRNA distribution (see Figure S6.16C). This was attributed to the fact that *ptsG* is known to be under regulatory control [439]. To demonstrate the utility of the regulated theory (Eqs. S6.50–6.55), we used the mean and Fano factor calculated from the experimental distribution to constrain k_{on} and k_{off}.

We begin by computing $\langle m \rangle_1$ (according to Eqn. 6.77) and noting:

$$\langle m \rangle_1 = \frac{k_{\text{on}}}{k_{\text{on}} + k_{\text{off}}} \frac{k_t}{k_d} \tag{6.81}$$

We can solve this for α, the ratio of k_{off} to k_{on}, as:

$$\alpha = \frac{k_t}{k_d \langle m \rangle_1} - 1 \tag{6.82}$$

From the experimental distribution we can also compute the Fano factor;

by subtracting off the contributions associated with gene duplication and RNAP variability we can arrive at the Fano factor contribution associated with regulation (see Eqns. 7 and 8 in the main manuscript):

$$\begin{aligned}\text{Fano}_{\text{reg}} &\approx \frac{\text{Var}[m]}{\langle m\rangle} - 1 + \langle m\rangle - \frac{\langle m\rangle}{10} \\ &- \frac{\langle m\rangle_1^2}{\langle m\rangle}\left[1 + 3f + \frac{8e^{-fk_dt_D} - e^{-2fk_dt_D} - 7}{2k_dt_D}\right] \\ &= \frac{k_tk_{\text{off}}}{(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + k_d)}\end{aligned} \tag{6.83}$$

which we can then solve for k_{on} and k_{off}:

$$\begin{aligned}k_{\text{on}} &= \frac{1}{1+\alpha}\left[\frac{k_t\alpha}{(1+\alpha)\text{Fano}_{\text{reg}}} - k_d\right] \\ k_{\text{off}} &= \alpha k_{\text{on}}\end{aligned} \tag{6.84}$$

Because of the constraints on k_{on} and k_{off}, the fitting problem reduced to a single dimensional scan over values for k_t, and therefore the computational time required to fit the distribution was greatly reduced. Simulations demonstrate that $k_t = 3.95 \pm 0.1 min^{-1}$ best fit the distribution (Figs. S6.16A–C). This parameter was found to be robust to bin size for the data (Fig. S6.16C).

Uncertainty in the input data will make the solution to the fitting problem non-unique. In order to test this effect on the fit, we varied k_d within $\pm 1\sigma$ of the mean value reported and ran linear scans over reasonable k_t value to identify the optimal fit transcription rate and regulation parameters k_{on} and k_{off}. Various metrics comparing simulated distributions of mRNA compared to experimental distributions in Fig. S6.17. All metrics demonstrate that the optimal solution falls on a line in $k_t - k_d$ plane that corresponds to values

for k_{on} and k_{off} that are indistinguishable, within uncertainty, of the values obtained from fitting to the average k_d (p=0.256 for k_{on} and p=0.892 for k_{off} by t-test; see Fig. 6.17).

6.5.10 Results of Simulations Including Regulation and RNAP Variability

Comparisons of simulations including regulated gene expression with $k_{\text{on}} = k_{\text{off}} = 0.2/min$ with the theories are shown in Fig. S6.18. In these simulations k_t was taken to be 2.52min^{-1} to maintain the same averages seen as in the constitutive expression, so that resulting noise can be compared. Again, agreement is nearly exact.

Contributions of extrinsic noise to the total noise was approximated by including RNAP variability in simulations. The average RNAP were taken to be 2500 and 5500 for cells doubling in 70 and 40 minutes, respectively [436]. RNAP distributions were modelled as Γ-distributions with the shape parameter of 10 and scale parameters of 250 (40 min) or 550 (70 min). A total of 2000 cells were simulated in each case and for each cell a single RNAP count was sampled from the respective Γ-distribution and held constant for ten cell cycles. Simulation results for the average and Fano factor are compared to the analytical theory assuming now that the variation affects k_t (Eqn. 8 in the main manuscript or Eqn. S6.71) in Figure S6.19. This type of noise is accurately captured by the theory.

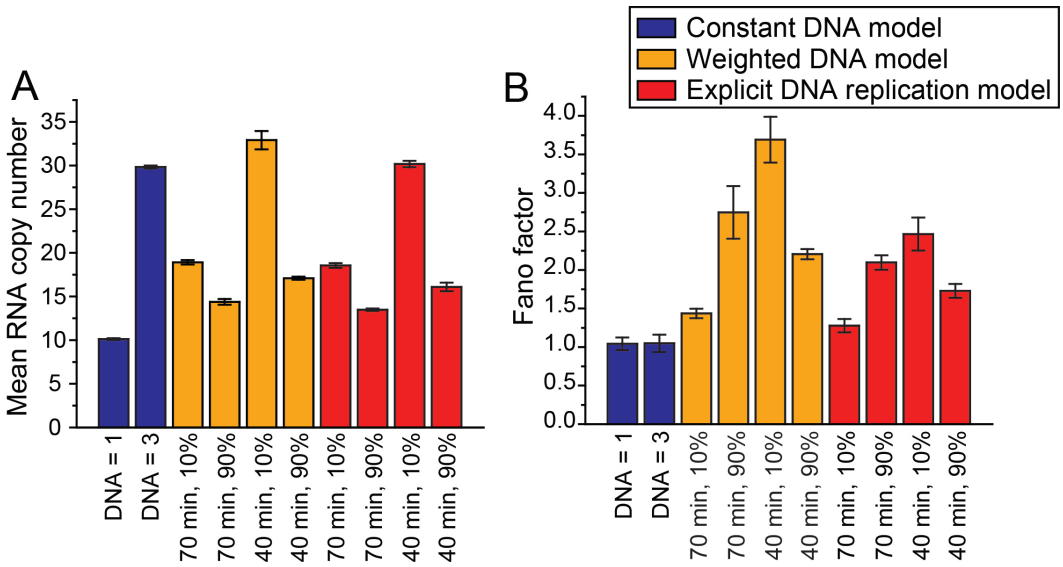


Figure 6.6: **Division Time Contribution.** Comparison of various average mRNA levels (A) and their associated Fano factors (B) for different treatments of gene copy number in stochastic simulations. The "Constant DNA model" assumes that there is only one gene copy number and all cells in the population have that number over all time. The "Weighted DNA model" is equivalent to the time-independent theory, in that each cell is considered to have either a high or a low count of the gene based on the fraction of time after gene replication, f, and assumed to have that copy nubmer for all time. The "Explicit DNA replication model" is that of the time-dependent theory, where the gene is duplicated during the simulation and the mRNA is allowed to relax to the new steady-state. Simulations with genes at different locations (10 or 90% of the distance from the origin to terminus) at two doubling times are considered. The noise observed in the explicit replication model is consistently lower than that in the weighted DNA model.

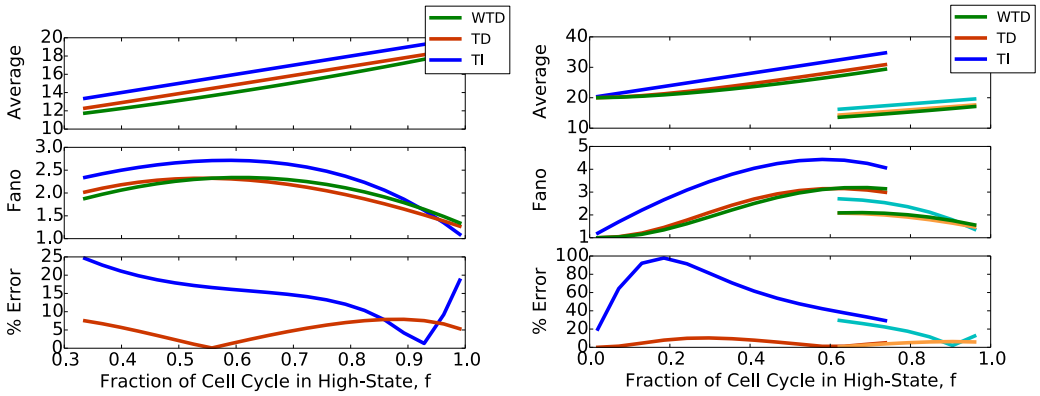
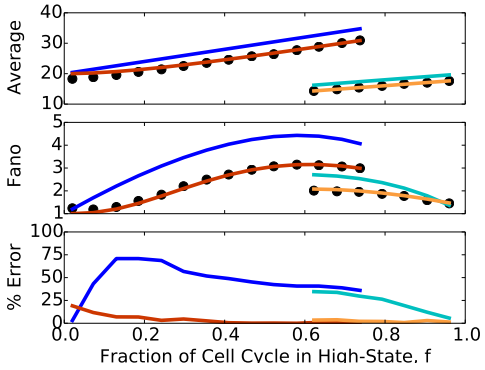
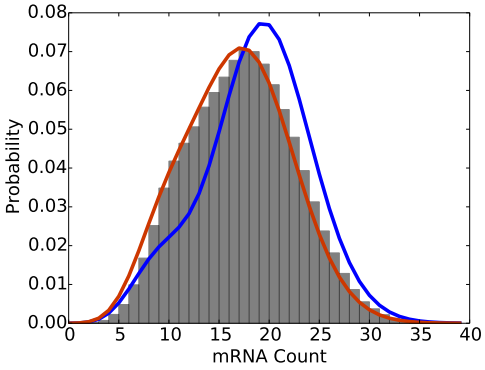


Figure 6.7: **Cell–Age Weighted Results.** Comparison of average mRNA and Fano factor predicted by theories with (orange/light orange lines) and without (blue/cyan lines) accounting for time dependent mRNA to the theory that weights the results with exponentially distributed cell ages (green) for cells doubling in (a) 70 minute and (b) 40 minutes. The form of the weighted time-dependent theory (WTD) is based on Eq. S6.36. The bottom plot shows errors of the TD and TI theories relative to the exponentially weighted TD theory (green lines), which are generally below 8%.



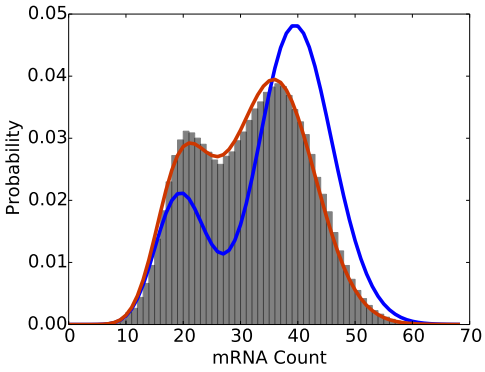
(a)



(b)

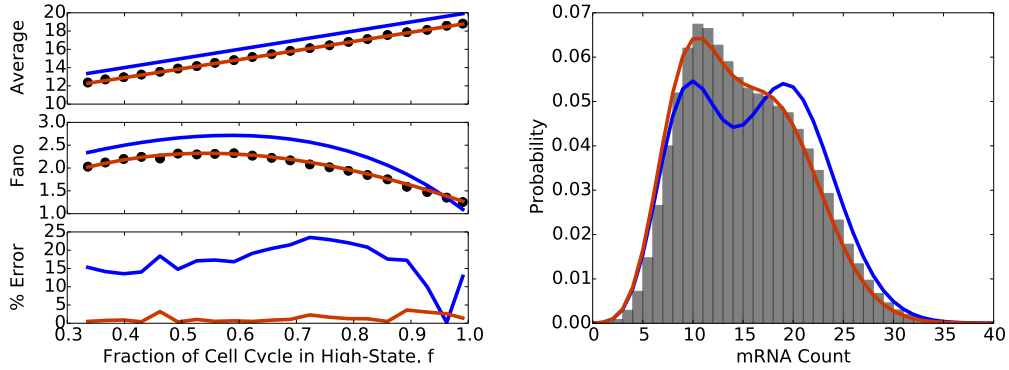
Gene locus	40 minute cell cycle			
(% from *ori*)	N = 1	N = 2	N = 4	Mean
1		0.261	0.739	3.478
10		0.3625	0.6375	3.275
20		0.475	0.525	3.05
30		0.5875	0.4125	2.825
40		0.7	0.3	2.6
50		0.8125	0.1875	2.375
60		0.925	0.075	2.15
70	0.0375	0.9625		1.9625
80	0.15	0.85		1.85
90	0.2625	0.7375		1.7375
100	0.375	0.625		1.625

(c)



(d)

Figure 6.8: **40 Minute Doubling Time.** A) Comparison of average mRNA and Fano factor predicted with (orange/light orange lines) and without (blue/cyan lines) accounting for time-dependent mRNA relaxation to results from exact simulations (points). Darker orange and blue lines represent genes that are duplicated during the cell cycle when the replication of that genome is initiated, and therefore have either 2 or 4 gene copies. Lighter orange and cyan lines represent genes that are duplicated in the cell cycle following the one in which the replication was initiated, and therefore either 1 or 2 gene copies exist. The time dependent theory shows nearly exact agreement. Comparisons of the mRNA distribution from simulation (gray histogram) to theories with (orange lines) and without (blue lines) time-dependence demonstrates the advantange of considering the mRNA relaxation for genes that spend (B) 27.5% and (D) 62.5% of their time in the high state.

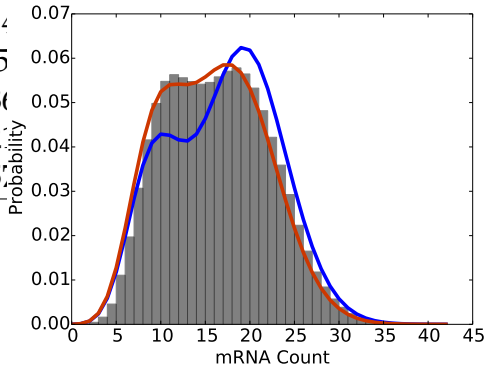


(a)

(b)

Gene locus	70 minute cell cycle		
(% from *ori*)	N = 1	N = 2	Mean
1	0.0064	0.9936	1.9936
10	0.064	0.936	1.936
20	0.129	0.871	1.871
30	0.193	0.807	1.807
40	0.257	0.743	1.743
50	0.312	0.679	1.67
60	0.386	0.614	1.61
70	0.45	0.55	1.55
80	0.514	0.486	1.48
90	0.579	0.421	1.42
100	0.643	0.357	1.35

(c)



(d)

Figure 6.9: **70 Minute Doubling Time.** A) Comparison of average mRNA and Fano factor predicted with (orange lines) and without (blue lines) accounting for time-dependent mRNA relaxation to results from exact simulations (points). The time-dependent theory shows nearly exact agreement in all cases. When comparing numerically computed distributions for genes that spend (B) 61% and (D) 74.3% of the cell cycle in the high state, to simulated distributions (gray histograms) it becomes apparant that including time-dependence (orange lines) better captures both qualitatively and quantitatively the data than does the time-independent theory (blue lines).

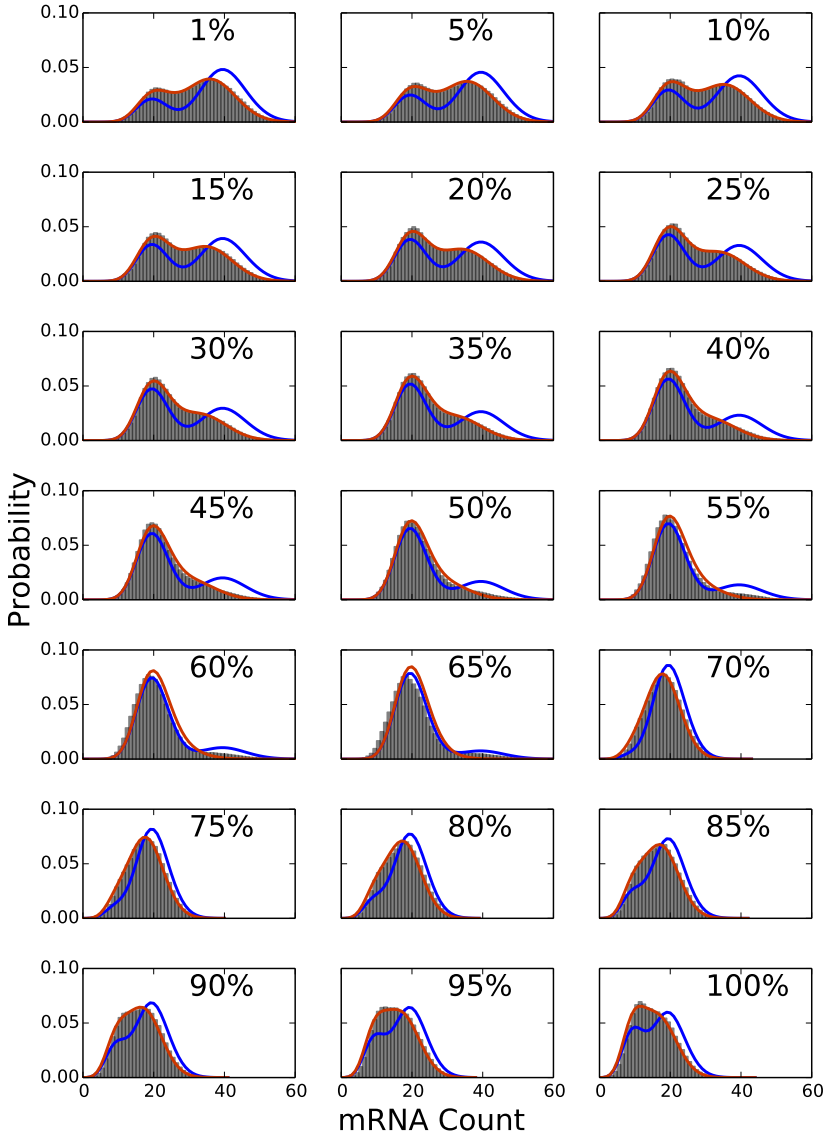


Figure 6.10: **40 Min Doubling Time Distributions.** Results for distributions computed via theory for a cell doubling every 40 minutes for a genes located at the indicated positions between the origin and the terminus. Distributions computed with the analytical theory (orange lines) nearly exactly represent simulations (gray distributions) whereas distributions computed via the time-indepenent model (blue lines) often qualitatively predict strong bimodal behavior, where none should exist. In all cases, the time-dependent theory is superior as demonstrated in Figure 6.12. However, as discussed in the main text, the comparison becomes worst between 50-60% of the way along the genome, due to inadequite time to relaxation to the high-state.

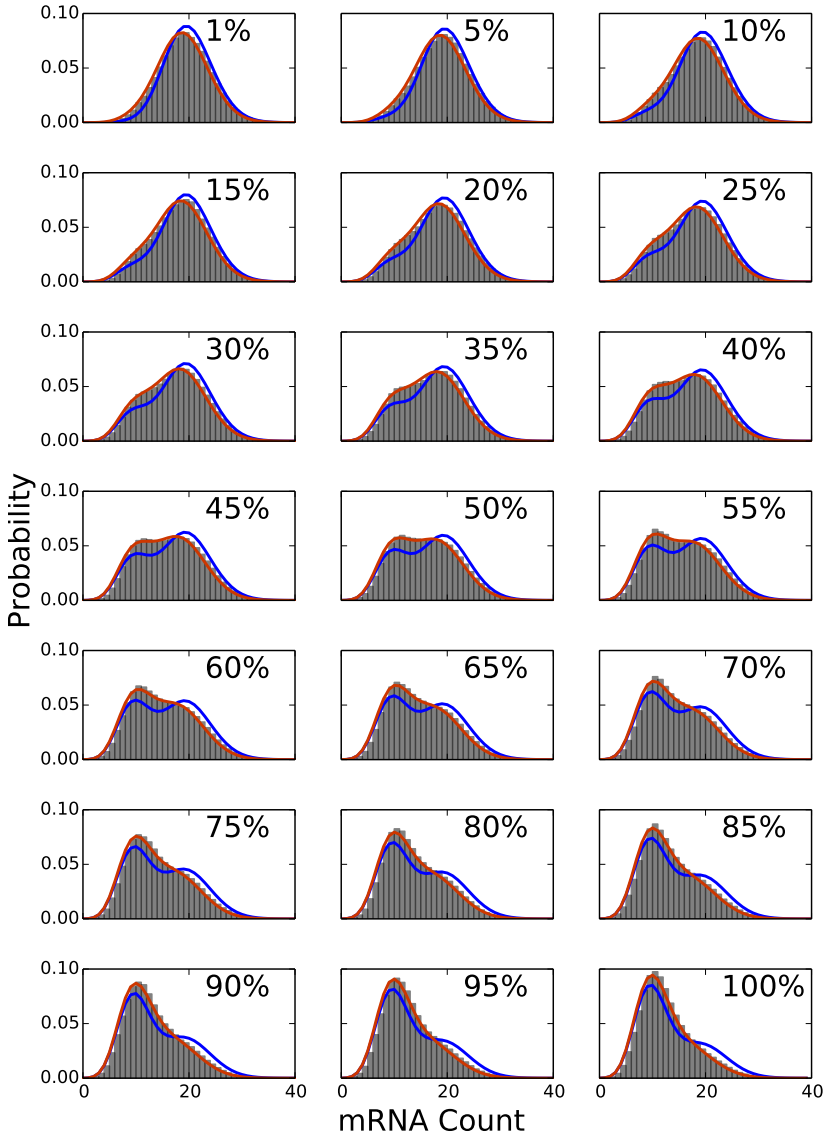
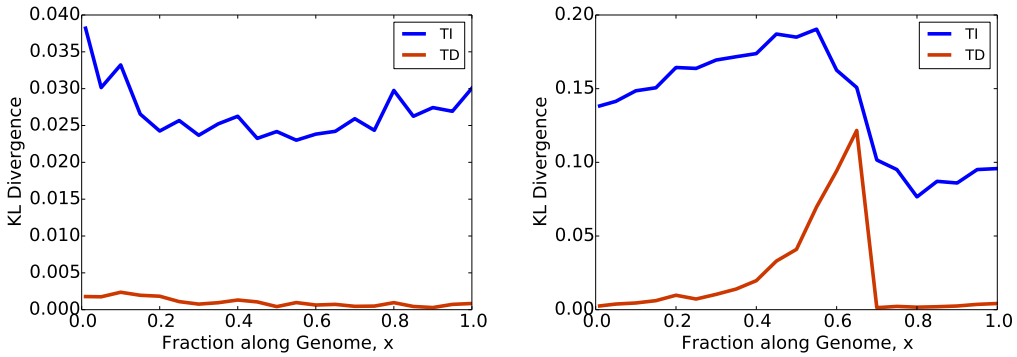


Figure 6.11: **70 Min Doubling Time Distributions.** Results for distributions computed via theory for a cell doubling every 70 minutes for a genes located at the indicated positions between the origin and the terminus. Distributions computed with the analytical theory (orange lines) nearly exactly represent simulations (gray distributions) whereas distributions computed via the time-indepenent model (blue lines) often qualitatively predict strong bimodal behavior, where none should exist. In all cases, the time-dependent theory is superior as demonstrated in Figure 6.12.



(a) (b)

Figure 6.12: **Goodness of Fit.** Kullback-Leibler divergence of theory from simulation computed for doubling times of (a) 70 and (b) 40 minutes. The theory incorporating time-dependent mRNA dynamics (orange lines) better captures the mRNA distribution than does the static theory (blue lines). The rapid rise in divergence in the 40 minute doubling time comparison is due to the fact that mRNA from genes that duplicate close to division time does not have enough time to relax to the new steady-state prior to division, thus starting the next cell cycle away from steady-state (see Fig. 6.13 description in the main text).

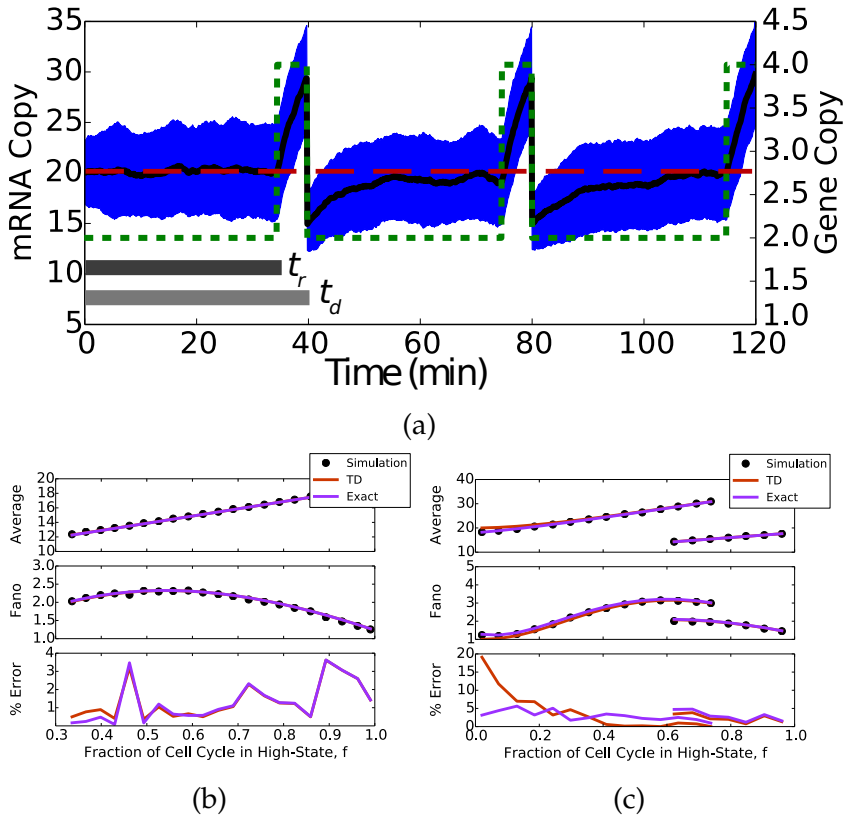


Figure 6.13: **Deviation Near Division.** A) A schematic composed of 200 simulation replicates showing the progress of the average mRNA (black line) levels before and after a gene duplication event (green dotted line). The area encompassing the average $\pm 1\sigma$ (blue). As can be seen, replication is followed by relaxation of the mRNA from an initially low level but does not relax to a new steady-state level prior to cell division, and therefore the cells begin their next cell cycle with a non-steady-state mRNA distribution. Only about half-way through the next cell cycle does the mRNA approach steady-state (red dashed line). In a case such as this, the TD theory put out in the paper will deviate, as it was derived with the assumption that the mRNA reaches steady state before division. A modified TD theory lifts this assumption allowing for more accurate estimation of the average, variance, Fano factor, and mRNA distributions (Eqs. S35-37). The doubling time (t_D) was taken to be 40 minutes, the total DNA replication time was taken to be 45 minutes, the gene was positioned 55% of the way from the origin to the terminus ($t_r \approx 35$ minutes), the transcription rate k_t was 1.26 min^{-1} and the degradation rate k_d was 0.126min^{-1}. The assumption that the mRNA level relaxes to steady-state prior to cell division can be lifted and "Exact" equations can be derived (purple lines; Eqs. S6.45-6.46) that better capture the Fano factor, especially for genes that replicate close to cell division (f <0.1), as demonstrated for the 70 minute (B) and 40 minute (C) cases.

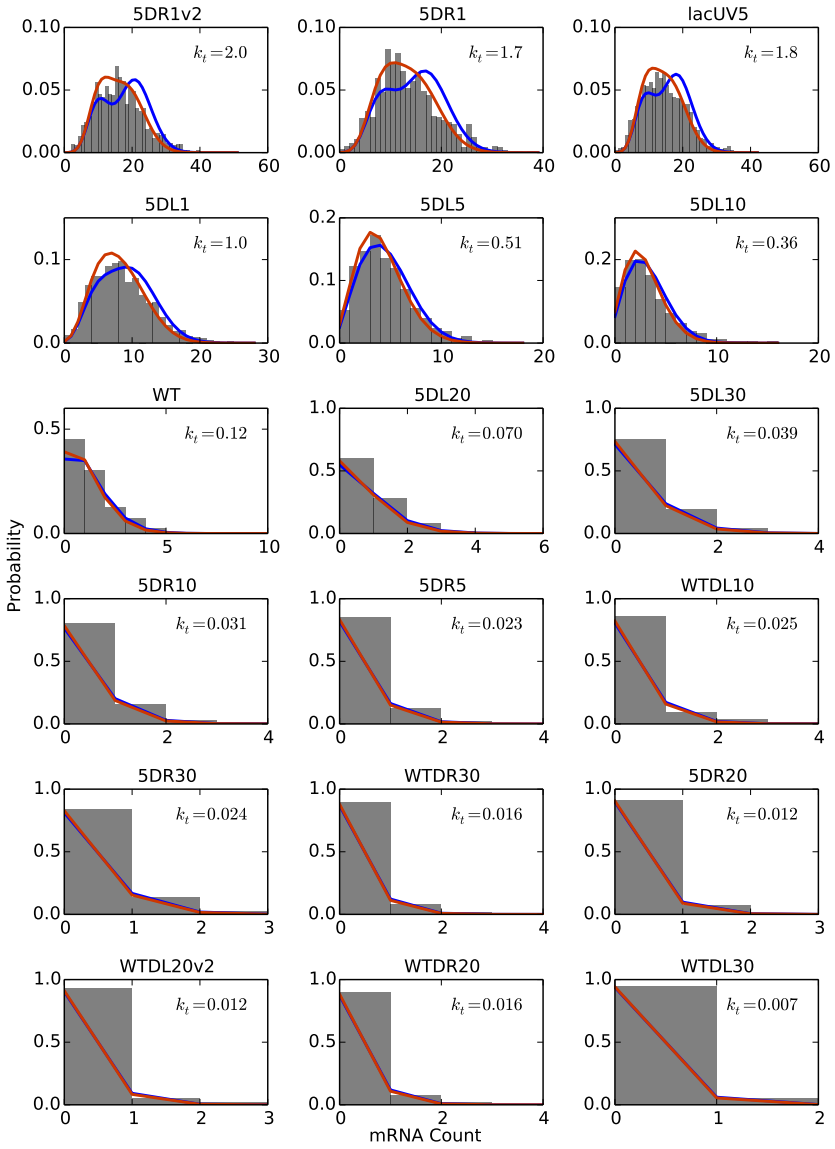
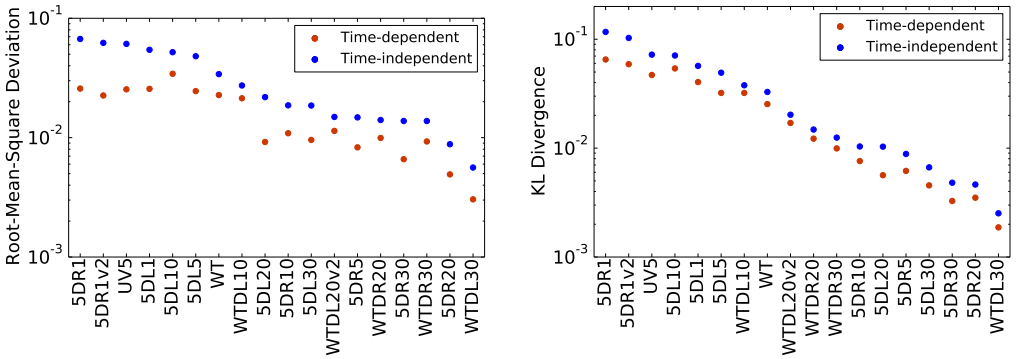


Figure 6.14: **Comparison to Experimental Distributions.** Comparison of time-dependent (orange) and time-independent (blue) theories for the mRNA distribution to experimental data of Jones et al. [49]. Theoretical curves are computed taking half-life and doubling times of 5.5 and 40 minutes, respectively, for a gene that is in the high-state for 2/3 of the cell cycle. The mean from a single gene copy was computed as disscussed in Section S6.5.9; the associated transcription rate is shown in the figure in units of per-minute.



(a) (b)

Figure 6.15: **Distribution Comparisons.** A) Mean squared deviation (MSD) computed (via Eq. S6.80) between the predicted and experimental mRNA distributions of Fig. S6.14. B) KL-divergence computed (via Eq. S6.76) on the same data. Both metrics demonstrate that the TD theory better represents the data than does the TI theory. Smaller MSD or KL indicates better agreement.

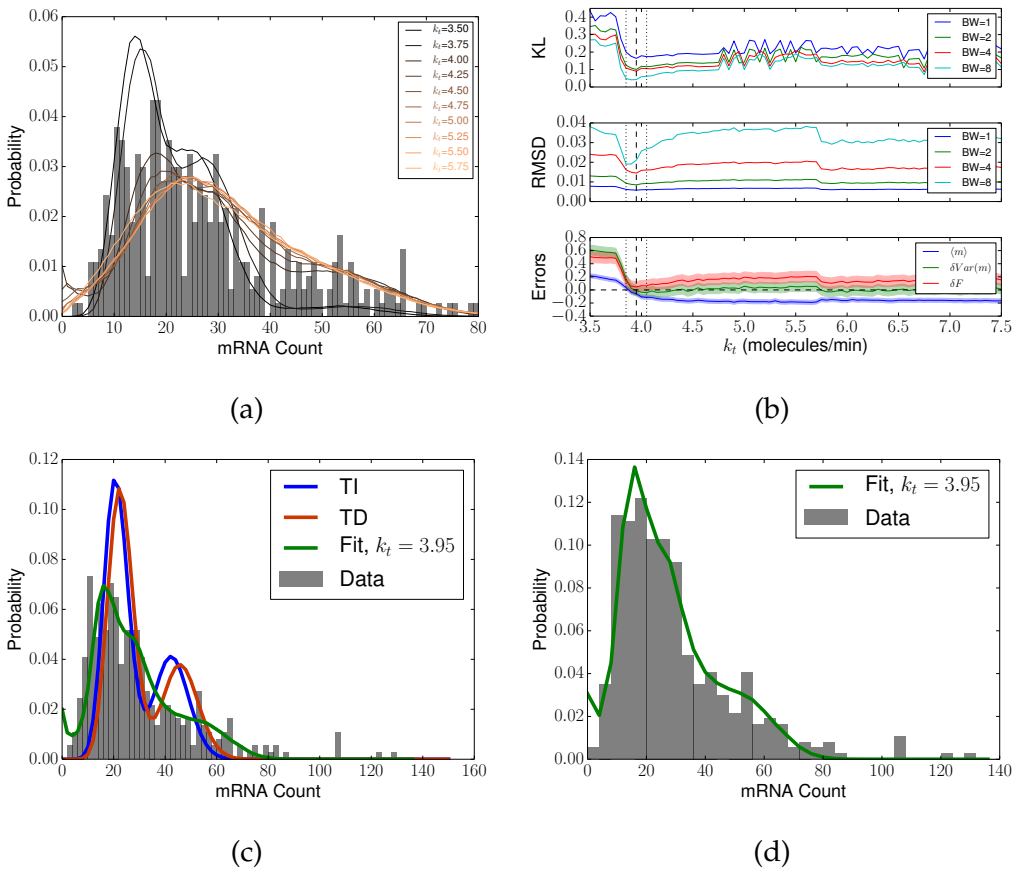


Figure 6.16: **Fit to *ptsG* Distribution.** Fitting of the experimental distribution (of two pooled replicates) for *ptsG* via simulations while constraining k_{on} and k_{off} via the regulated theory (Eqs. S6.50–6.55). A) By varying k_t distributions were simulated. B) Distributions were compared with experimental data using various metrics such as KL divergence (top) and RMSD (mid) at various binning widths (BW), from which the most optimal k_t was chosen (dashed vertical line; ± 1 SEM dotted vertical lines), that also minimizes the error in the mean, variance and Fano factor. The corresponding regulation rates are $k_{\text{on}} = 0.023 min^{-1}$ and $k_{\text{off}} = 0.0084 min^{-1}$. This model significantly outperforms the constitutive theory as shown for comparisons with bin widths of C) 2 and D) 4. The noise of the observed distribution can only be caputed with this model as *ptsG* is a highly regulated gene [439]. Using the analytical theory, only 1 parameter was varied; therefore the effort expended on fitting was significantly reduced. Other model parameters include a half-life of 2.8 minutes for a gene located 25% from the *ori*, corresponding to $f \approx 0.35$ [156].

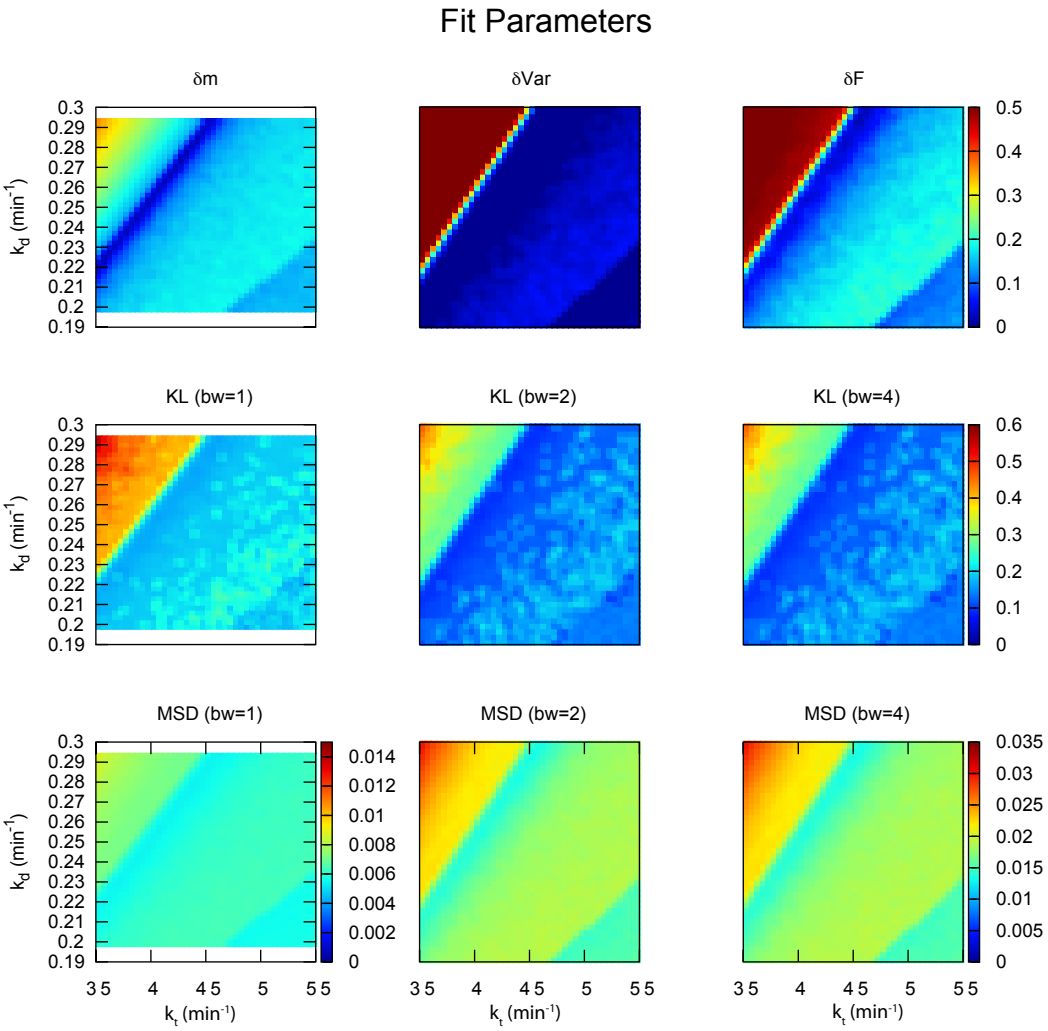
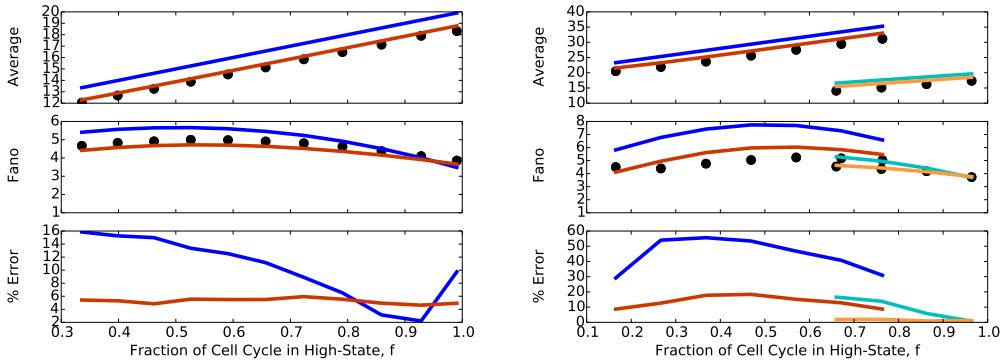


Figure 6.17: **Fitting Statistics.** Various measures of fit agreement between simulated and experimental distributions as k_d is varied within 1σ of the average plotted versus the free parameter k_t. (Top Row) Absolute value of the relative error in the mean δm, variance δVar and Fano factor δF for the simulated parameter sets. (Middle Row) Computed Kullback Leibler-divergence for various histogram binning widths. (Bottom Row) Computed mean–squared–deviation for various histogram binning widths.



(a) (b)

Figure 6.18: **Approximated Regulated Noise.** Comparison of average mRNA and Fano factor predicted by theories with (orange/light orange lines) and without (blue/cyan lines) accounting for time dependent mRNA to simulation results (points) for (a) 70 minute and (b) 40 minute doubling times. The form of the time-dependent theory is based on the approximate corrections of Eq. S6.55.

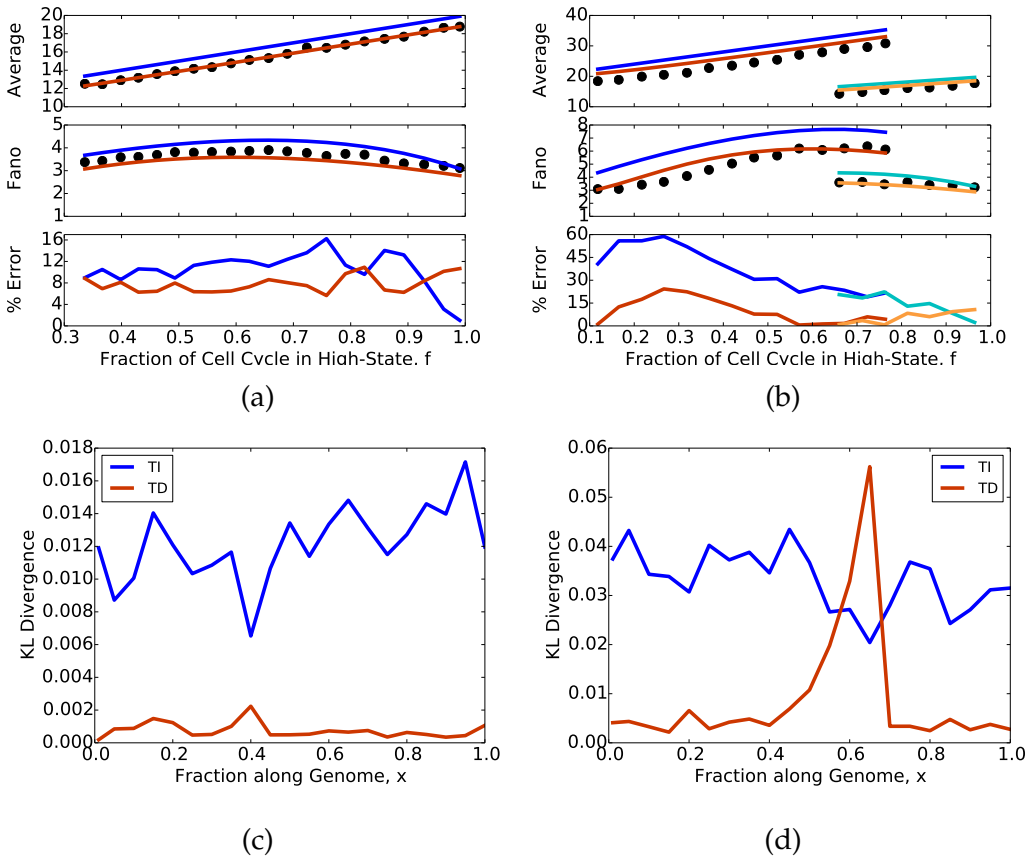


Figure 6.19: **RNAP Noise.** A comparison of theory with simulations where extrinsic noise is entirely approximated by variations in the RNAP number, and consequently variation in the apparent transcription rate k_t. (A) 70 minute and (B) 40 minute doubling times. The Kullback-Leibler divergence comparing numerically computed distributions with simulated distributions are shown for (B) 70 minute and (D) 40 minute doubling times demonstrating that incorporation of the time-dependent mRNA relaxation is required even when considering extrinsic effects such as RNAP fluctuations. The time-dependent theory is calculated using the approximate solution Eq. S6.71. The time-independent theory is based on Eqs. S33 and S54 of Jones et al. [49].

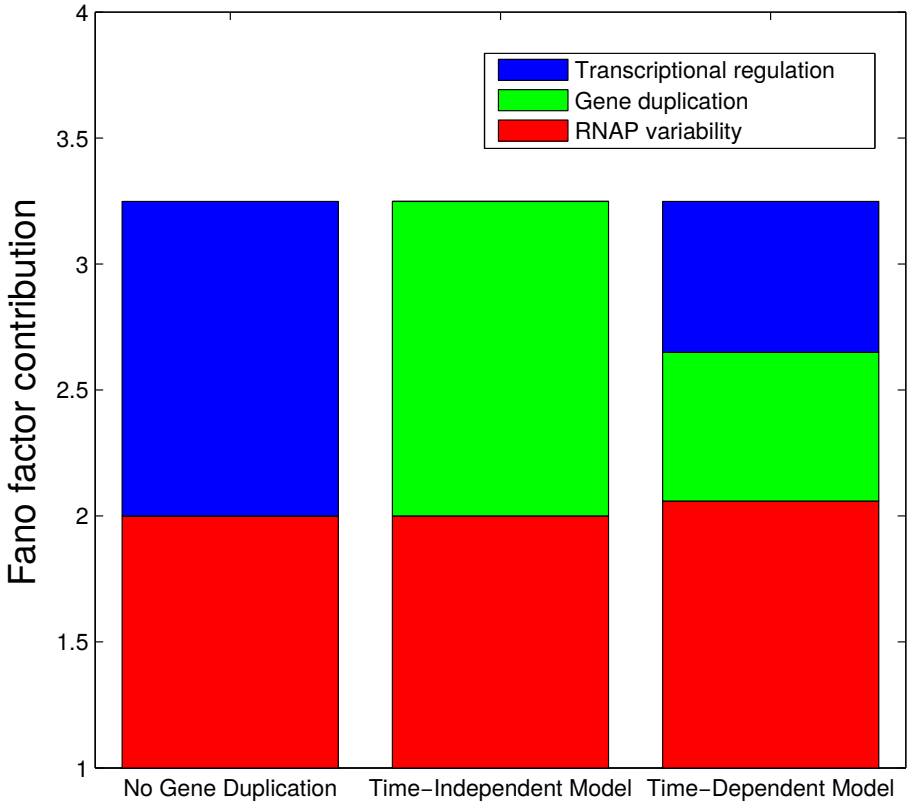


Figure 6.20: **Noise Contributions.** Noise contributions assessed using different models of mRNA noise. In each case it is assumed $t_D = 40$ min, $\langle m \rangle = 10$, $f = 0.35$, $k_d = 0.126$ min^{-1}, and Fano$[m] \approx 3.25$. In the left-most bar, the noise is assumed to originate entirely from RNAP variability or transcriptional regulation. In the central bar, noise contributions are computed according to the time-independent theory. In this case RNAP variability and gene duplication alone completely account for the observed Fano factor; in turn, transcriptional regulation appears not to contribute. In the right-most bar, noise contributions are computed according to the time-dependent theory. In this case we see that the time-independent theory overestimates gene duplication-associated noise, obscurring the fact that some transcriptional regulation is taking place, although not as much as might have been suspected had gene replication not been accounted for.

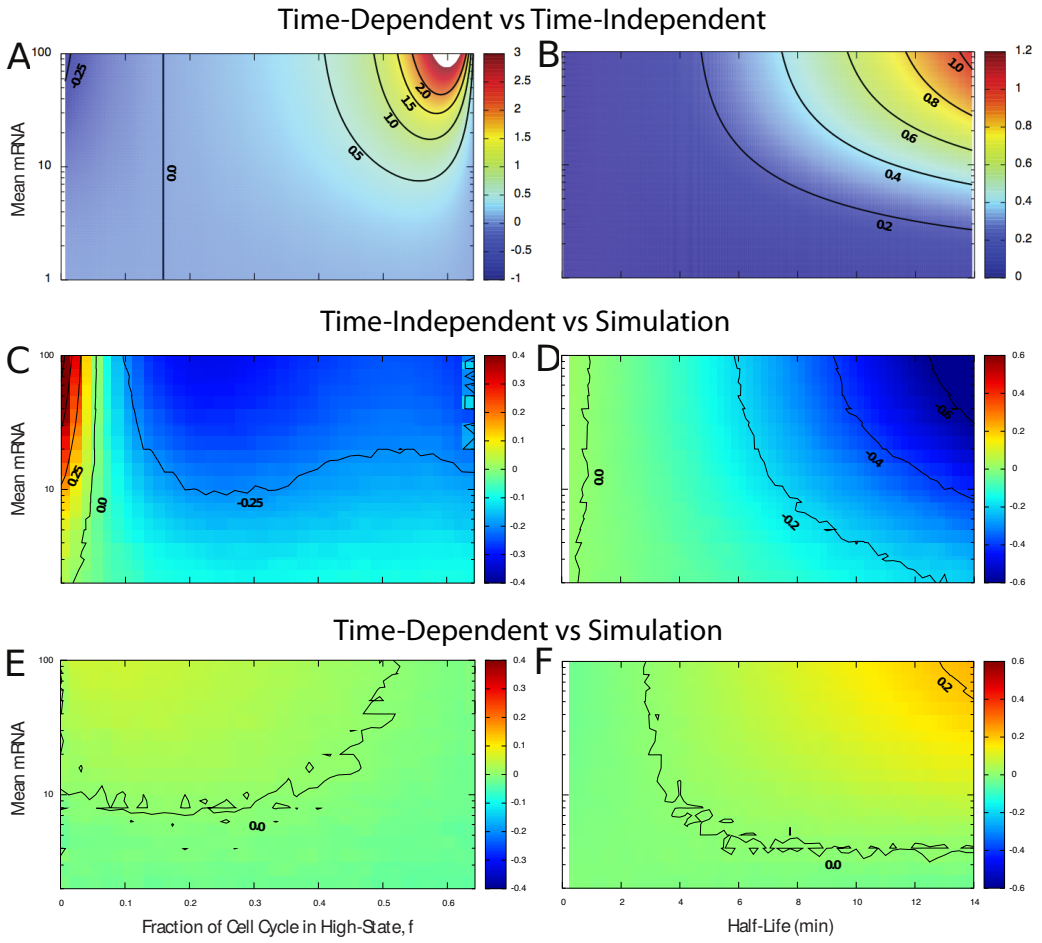


Figure 6.21: **Deviation of Theories and Simulations.** (A) and (B) as in Fig. 4 where the error is $(F_{TI} - F_{TD})/F_{TI}$. Comparison of the Fano factor computed via theory to stochastic simulations can be seen for the TI ($(F_{Sim} - F_{TI})/F_{Sim}$; C & D) and TD ($(F_{Sim} - F_{TD})/F_{Sim}$; E & F) expressions. Simulations were averages of 1000 indpendent cell lineages each growing for 10 generations. Contour lines are for the indicated values, and are not smooth due to the variation in the data due to limited sampling. The average deviation of the TD theory is generally less than 20% over the ranges studied, while the TI can deviate by over 60% in the same ranges. The error in E & F can be reduced to zero within numerical uncertainty and sampling error by using Eqs. S35-37 as opposed to Eq. S28.

Chapter 7

Parameteric Studies of Metabolic Cooperativity in *Escherichia coli* Colonies: Strain and Geometric Confinement Effects

Characterizing the complex spatial and temporal interactions among cells in a biological system (*i.e.* bacterial colony, microbiome, tissue, *etc.*) remains a challenge. Metabolic cooperativity in these systems can arise due to the subtle interplay between microenvironmental conditions and the cells' regulatory machinery, often involving cascades of intra- and extracellular signalling molecules. In the simplest of cases, as demonstrated in a recent study of the model organism *Escherichia coli*, metabolic cross-feeding can arise in monoclonal colonies of bacteria driven merely by spatial heterogeneity in the availability of growth substrates; namely, acetate, glucose and oxygen. Another recent study demonstrated that even closely related *E. coli* strains evolved different glucose utilization and acetate production capabilities, hinting at the possibility of subtle differences in metabolic cooperativity and the resulting growth behavior of these organisms. Taking a first step towards understanding the complex spatio-temporal interactions

The contents of this chapter are based in part on work previously published as Joseph R. Peterson, John A. Cole and Zaida Luthey-Schulten. "Parameteric Studies of Metabolic Cooperativity in *Escherichia coli* Colonies: Strain and Geometric Confinement Effects," *PLoS ONE*, 12(8):e0182570. [59]. Specifically, J.A.C. provided the code and helped with analysis throughout the work.

within microbial populations, we performed a parametric study of *E. coli* growth on an agar substrate and probed the dependence of colony behavior on: 1) strain-specific metabolic characteristics, and 2) the geometry of the underlying substrate. To do so, we employed a recently developed multiscale technique named 3D dynamic flux balance analysis which couples reaction-diffusion simulations with iterative steady-state metabolic modeling. Key measures examined include colony growth rate and shape (height vs. width), metabolite production/consumption and concentration profiles, and the emergence of metabolic cooperativity and the fractions of cell phenotypes. Five closely related strains of *E. coli*, which exhibit large variation in glucose consumption and organic acid production potential, were studied. The onset of metabolic cooperativity was found to vary substantially between these five strains by up to 10 hours and the relative fraction of acetate utilizing cells within the colonies varied by a factor of two. Additionally, growth with six different geometries designed to mimic those that might be found in a laboratory, a microfluidic device, and inside a living organism were considered. Geometries were found to have complex, often nonlinear effects on colony growth and cross-feeding with "hard" features resulting in larger effect than "soft" features. These results demonstrate that strain-specific features and spatial constraints imposed by the growth substrate can have significant effects even for microbial populations as simple as isogenic *E. coli* colonies.

7.1 Introduction

Metabolic competition and cooperativity are ubiquitous in nature with recent research reaffirming the old adage: location is everything. Whether it be the division of labor among the cells in an animal body or the complex chemical warfare among a soup of bacteria competing for limited resources, spatial and temporal variation are key factors which must be understood. Interactions in microbial communities, which often comprise tens to hundreds of metabolically distinct species [440], are of particular interest in areas ranging from human health [441] to ecology of the world's nutrient cycles [442, 443]. These communities form complex networks of cooperative and competitive interactions that ultimately determine the population's dynamics, steady-states, and robustness to change [444]. For example, it is now known that among people with chronic bowel diseases, the composition of the gut microbiota can vary considerably relative to the "healthy" patient [445, 446]. More generally, the potential for emergent metabolic cooperativity via a wide array of organic acids, inorganic molecules, salts, and purine/pyrimidines have been predicted in two-member bacterial communities [447]. Underlying the stability and structure of these populations are a complex network of metabolic and physical interactions that vary both spatially and temporally [448]; thus, part of what is needed to understand how these populations behave is an understanding of the metabolism of community members growing alone and in concert with their neighbors.

A number of computational systems biology techniques have been devel-

oped for understanding the growth and metabolic requirements of different microbes. Chief among these is a family of methods based on flux balance analysis (FBA) [69]. FBA describes the steady-state growth and fluxes within an organism's metabolic network subject to internal constraints (*e.g.* reaction upper bounds proceeding from the finite copy numbers of catalytic enzymes and their turnover rates) and external constraints (*e.g.* limited availability of a molecule needed for growth). Several FBA methods designed to examine communities of microorganisms have been developed including community FBA (cFBA), OptCom, dynamic multi-scale FBA (dmsFBA) and population FBA [56, 266, 267, 447, 449]. cFBA compartmentalizes different organisms but allows them to compete for and exchange metabolites through an extracellular compartment [267]. OptCom uses a similar technique, however it allows for a community objective to be defined, which enables users to identify optimal engineering strategies for a community [266]. dmsFBA relaxes the common steady-state assumption in order to simulate how cross-feeding evolves with time [447]. Population FBA simulates metabolic phenotypes in populations of microbes that arise due to capacity constraints that arise from stochastic gene expression [56]. For an excellent review of these and other community methods and studies, see [450]. All of these methods are limited in that they treat the entire population as being "well-stirred", that is, seeing the same chemical and spatial environment. This may be a reasonable approximation for some problems (e.g. bioreactors), but spatial heterogeneity is innate to many important scenarios (*e.g.* biofilms, microbiome).

A recently developed multi-scale systems biology technique provides

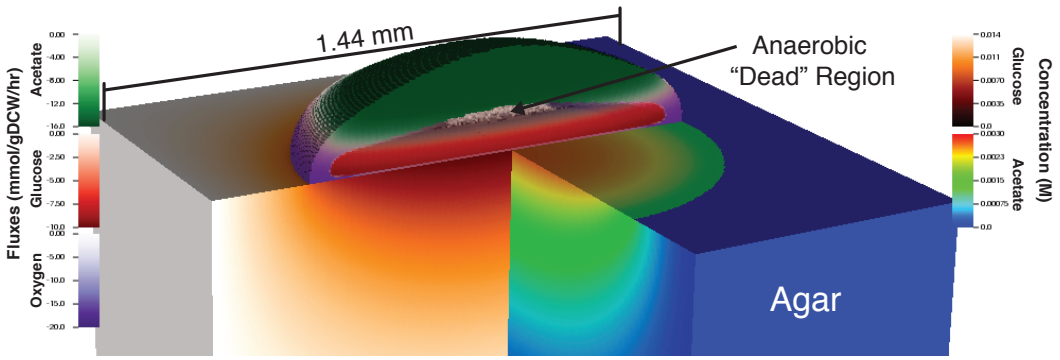


Figure 7.1: **Metabolic Crossfeeding in *E. coli*.** A visualization of a 3DdFBA simulation of metabolic cooperativity in *E. coli* K-12 MG1655 growing on a 2.5% glucose agar substrate (described previously [58]). The image depicts metabolic cooperativity between aerobic (purple) and anaerobic (red) glucose utilizing cells, which generate acetate that is utilized aerobically by glucose starved cells (green). Complete substrate utilization at the top and bottom of the colony create an anoxic dead zone containing essentially dormant cells. Color bars on the left represent metabolic fluxes of the indicated metabolites while those on the right indicate concentrations of glucose and acetate inside the agar. The image was visualized using the VisIt visualization software [451] with a custom plugin written to support 3DdFBA simulation.

a means to examine the spatial dependence of microbial growth. Named 3-dimensional dynamic flux balance analysis (3DdFBA) [58], this method solves a reaction-diffusion equation on a lattice, allowing chemicals to diffuse in and among various phases (*i.e.* solid, liquid, gas, cell, *etc.*) and be taken up by cells living in a given lattice site, effectively converting chemicals into biomass. The method has been formulated with both stochastic [155, 452] and continuous [58] descriptions of the reaction and diffusion of the chemical species. The stochastic version of the method was applied to small colonies ($\sim$100 cells) of *E. coli* growing in a micro-aerobic environment [155, 452]. These simulations suggested that metabolic cooperativity could arise in an isogenic *E. coli* population, which in turn prompted the development of a con-

tinuous version capable of simulating macroscopic colonies [58]. Laboratory scale simulations of *E. coli* colonies growing on 2.5% glucose minimal media predicted that after sufficient time the population would fractionate into glucose-utilizing cells that ferment acetate, and acetate-utilizing cells which are starved of glucose (Figure 7.1). A number of other spatially-resolved FBA methods have also been developed, and used to study mutualism and competition in two dimensional multispecies communities (COMETS; [453]) as well as chronic wound biofilm homoeostasis (DFBALab; [454, 455]).

These spatially-resolved methods have the potential to add new relevance to computational biology by bridging temporal and spatial scales that are currently difficult or impossible to study experimentally. As a demonstration of their utility for complex 3D geometries, we performed parametric studies of metabolic cross-feeding using 3DdFBA in: 1) colonies of closely related *E. coli* strains, and 2) colonies grown on agars with geometries other than merely flat surfaces. We also analyze the error resulting from the discretization of the reaction-diffusion equation and show that grids finer than 20 μm are required to ensure converged solutions. Our simulations demonstrate that 3DdFBA can predict significant effects on the dynamics of microbial populations that arise through subtle differences between strains. Finally, we show that while abrupt changes in the shape of a colony's substrate (*i.e.* the existence of a nearby wall of agar) can dramatically impact the colony's growth, more subtle curvatures (such as those that may arise within the gut) give rise to growth dynamics that are very similar to reference colonies grown on flat agar.

7.2 Methods

7.2.1 3DdFBA

Three dimensional dynamic flux balance analysis (3DdFBA) is only briefly reviewed here, as it was described rigorously previously [58]. The method couples a partial differential equation (PDE) description of chemical transport with flux balance analysis (FBA) [69]. Broadly speaking, the PDE represents the chemical species while FBA represents the cells. Metabolites are consumed and produced by a reaction-diffusion PDE:

$$\frac{\partial \vec{C}}{\partial t} = \mathbf{D}\nabla^2\vec{C} + R(\vec{C}) \tag{7.1}$$

where $\vec{C}$ is a vector containing the concentrations of metabolites, $\mathbf{D}$ encodes the diffusion rates of the metabolites and $R(\vec{C})$ encode the reactive fluxes of the species. R includes any reactions among chemical species, active and passive transport into and out of cell volume and, crucially, exchange fluxes computed via a local dynamic FBA (dFBA) [79] simulation (more precisely, fluxes are read from a table of solutions computed via FBA and the solution is used to compute uptake and efflux). Within this study, the reaction diffusion equation is solved on a 3 dimensional regular cubic lattice via a central finite difference scheme [80]. Dirichlet boundary conditions (constant value) are applied for all chemical species on the boundaries of the simulation volume. Gaseous species are allowed to diffuse anywhere in the simulation domain, while aqueous molecules (acetate, glucose, etc.) are limited to diffusion in

lattice sites containing agar and cell mass.

Cells, while being represented as a volume fraction (ϕ_i) on the lattice, are not diffused or actively transported (e.g. via chemotaxis) among lattice points. Rather, as cell growth occurs they are pushed isotropically into neighbouring lattice points (after some maximum volume fraction within the lattice site is achieved, namely $\sum_i \phi_i \geq 0.65$). Volume fraction is related the mass of cells as

$$\phi_i = \frac{m_i}{V \rho_i} \tag{7.2}$$

where m_i is the mass of a particular cell type in the lattice site, V is the volume of the lattice site, and $\rho_i = m_{i,cell}/V_{cell}$ is the density of a single cell with $m_{i,cell}$ and V_{cell} taken to be 258 fg and 1 fL [229], respectively. Cell mass grows exponentially at the rate set by the local dFBA as

$$\frac{dm_i}{dt} = v_{bm,i} m_i \tag{7.3}$$

where v_{bm} is the flux through the biomass equation. An absorbing boundary condition for the cell mass is applied to the boundaries of the simulation volume. Cell mass is prevented from penetrating into the agar substrate. The reaction term in Eq. 7.1 is coupled to the cell mass and the predicted uptake flux as $R(\vec{C}) = \vec{m} \cdot \vec{v}_C$ where $\vec{v}_C < \vec{v}_{C,max}$. The maximal uptake/secretion rate, $v_{C,max}$, is constrained assuming enzyme saturation effects (e.g. Michaelis-Menten kinetics for glucose uptake) and to prevent a chemical in a lattice site from becoming negative ($\vec{C} \geq 0$). The cell volume fraction couples

to the chemical concentrations by hindering diffusion (*e.g.* an attenuated diffusion rate computed according to a diffusion law that considers the local cell volume fraction [81]) and via the reaction term discussed above. Volume fractions (ϕ_i) for each cell phenotype are tracked at each lattice site, and a "regulation" function allow cells to transition between phenotypic states depending on the local concentrations of the chemical species. As an example, glucose utilizing cells can be converted into acetate utilizing cells if there is plentiful acetate but no glucose available at the lattice site for a significant amount of time. See [58] for more description of the "regulation" function.

Analysis, simulation and visualization codes to simplify the analysis of 3DdFBA simulation were created for this study. These codes can be found at `http://www.scs.illinois.edu/schulten/`. An analysis of the performance and memory requirements of the code can be found in SI Figure 1.

7.2.2 *E. coli* Strains

We selected five *E. coli* strains with curated genome scale metabolic models (GEMs) [311, 456, 457] for characterization: BL21, Crooks, MG1655, W and W3110. Chemostat experiments demonstrated variable glucose utilization efficiency and acetate production rates among these strains [457]. Additionally, some strains produce acetate during aerobic glucose growth, while others do not. Models for the *E. coli* strains (*i*JO1366, *i*B21_1397, *i*EcolC_1368, *i*WFL_1372 and *i*Y75_1357) were obtained from the BiGG Models database

version 1.3 [458]. All FBA simulations were performed using COBRApy [230]. For all simulations, the core *E. coli* biomass reaction [459] was used as the primary objective reaction.

When used unmodified, the models were incapable of predicting the correct growth rate and acetate production rate (neither aerobically nor anaerobically) when setting glucose uptake rates to those measured in the chemostat experiments. Therefore, the models were adjusted to minimize errors in acetate production and growth rates under both aerobic and anaerobic conditions. This was accomplished by fitting the maximum oxygen uptake rate and growth associated maintenance (ATP cost for cell growth) to experimental data. The fit parameters, glucose uptake rate, predicted acetate and growth rates, and the associated errors are shown in Table 7.1. In general, growth and aerobic acetate production rates could be fit with little error. The anaerobic acetate production rate was more difficult to capture; over the five strains we found an average error of 16.1%.

Table 7.1: ***E. coli* Growth Characteristics.** Growth characteristics of models of the *E. coli* strains examined in the current study. Results are shown for models [311, 456, 457] where maximal O_2 and growth associated ATP maintenance have been fitted to minimize deviation from experimental acetate production and growth rates.

Strain	**Growth Rate**[a]		$v_{glucose}$[b]		$v_{acetate}$[c]		v_{o2}[d]	**GAM**[e]	
	$+\mathbf{O_2}$[f]	$-\mathbf{O_2}$[g]	$+\mathbf{O_2}$	$-\mathbf{O_2}$	$+\mathbf{O_2}$	$-\mathbf{O_2}$		$+\mathbf{O_2}$	$-\mathbf{O_2}$
B21	0.76 (0%)[h]	0.29 (0%)	-8.0±0.3	-11.3±0.5	0.0 (0%)	9.17 (3.9%)	-15.5	60.25	50.45
Crooks	0.96 (0%)	0.77 (0%)	-12.5±0.5	-30.9±2.2	0.0 (0%)	25.2 (32%)	-33.5	121.45	59.75
MG1655	0.84 (15.2%)	0.46 (0%)	-9.5±0.3	-16.7±0.2	3.49 (0%)	13.27 (13.3%)	-13.9	60.25	50.45
W	0.97 (0%)	0.9 (0%)	-9.9±0.1	-27.2±1.4	0.0 (0%)	20.53 (2.6%)	-17.8	54.55	35.65
W3110	0.61 (0%)	0.52 (0%)	-6.7±0.1	-17.5±0.5	3.03 (2.7%)	13.63 (28.7%)	-7.5	36.65	41.65

[a]Optimal growth rate for cells grown in a chemostat in units of hr^{-1}. [b]Experimentally quantified maximal uptake rate for glucose in units of mmol/gDCW/hr [457]. [c]Maximal efflux rate for acetate in units of mmol/gDCW/hr. [d]Fitted maximal uptake rate for oxygen in units of mmol/gDCW/hr. [e]Fitted growth associated maintenance cost in units of mmolATP/gDCW. [f]$+O_2$ - Aerobic growth. [g]$-O_2$ - Anaerobic growth. [h]Values in parenthesis indicate percentage error compared to experimentally measured values after fitting.

Two phenotypes of *E. coli* were considered in the simulations: 1) those growing aerobically on glucose, and 2) those growing aerobically on acetate. Tables of FBA solutions (including uptake and efflux of key metabolites and growth rates) were generated for the phenotypes for all five *E. coli* strains. Because strain-specific acetate consumption data was not available, the maximal acetate uptake rate used in [58] was adopted. FBA tables were generate with 50 (160) divisions between 0 and the maximal glucose (oxygen) uptake rates.

7.2.3 Spatial Geometries

Spatial geometries that mimic engineered and natural growth environments were selected for analysis (Figure 7.2). A flat surface intended to mimic the standard agar in a Petri dish has been studied previously [58]. Additional geometries, namely the "plateau", "hole" and "wall", were designed to represent geometries that might be encountered in a microfluidic device (for instance in a microwell). For the plateau and hole geometries, two parameters were varied, namely the height (h) and width (w). For simplicity, the plateau and hole were taken to be square. For the wall geometry, the offset of the initial colony seed from the colony edge (o) is the only parameter. This parameter is of interest as the onset of metabolic cooperativity occurs after the colony has grown to some initial size and thus allows the investigation the effect of a confinement on onset. The final two geometries we studied include concave and convex surfaces that are designed to mimic biological systems such as the inside of an intestine, or the surface of a rough skin.

For the sake of simplicity, these geometries had uniform curvature defined through a radius parameter.

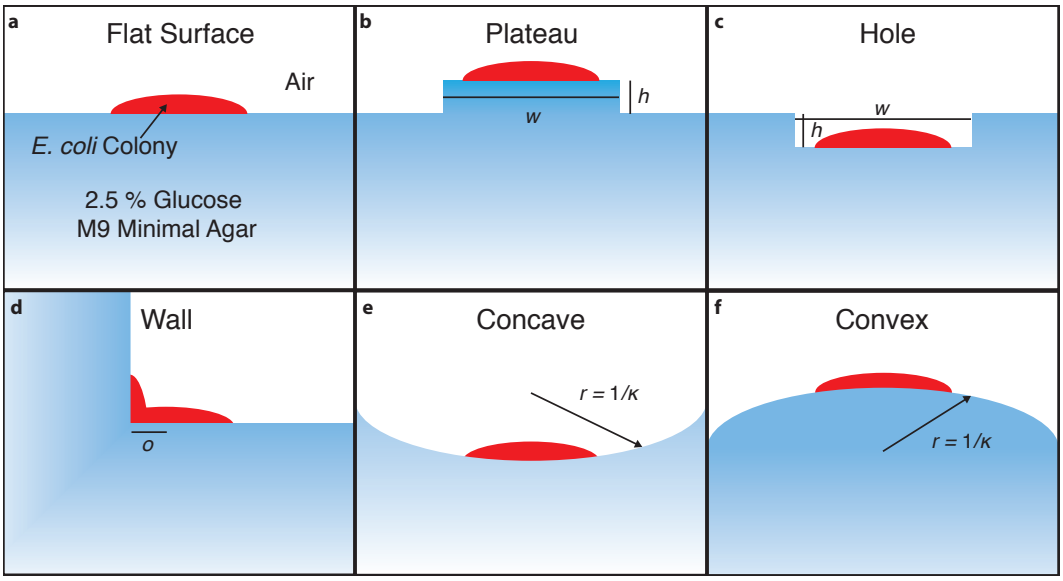


Figure 7.2: **Geometries Investigated.** An illustrative schematic showing the six geometries examined in this study. Various characteristic variables of the geometry are indicated with symbols. See the methods and results text for numerical values and descriptions of these variables.

7.3 Results & Discussion

7.3.1 Resolution Dependence of 3DdFBA Solution

The ability to resolve features within a 3DdFBA simulation depends on the resolution of the grid used to represent chemical concentrations and cell fractions. This resolution dependence is non-trivially dependent on diffusion and reaction rates, and on physical boundary conditions (*e.g.* the agar surface in the simulation). We examined how several features of interest (*e.g.* fluxes,

chemical concentrations, and cellular phenotypes) differ across a range of grid resolutions.

As a test system, we simulated the previously published *E. coli* K-12 MG1655 model [58] at eight grid resolutions ranging from 4.1 to 120 μm. Each colony simulation was seeded with the equivalent of a single cell's mass on the surface of the agar in the center of the simulation domain. Growth of the resulting colony was simulated for 40 hours to allow sufficient metabolic cooperativity to arise. Colonies grew to fill the agar and began to interact with the simulation boundary conditions at about 35 hours of growth. After this simulations began to exhibit boundary effects due to the use of fixed-concentration (Dirichlet) boundary conditions. Therefore, the resolution dependence was examined at the 30 hour time-point before any significant boundary effects arose.

We examine first the resolution dependence of the chemical concentration profiles within the colony; this particular feature is important in driving the partitioning of community members into different metabolic phenotypes during the simulation (Figure 7.3). Profiles taken through the colony's central axis show that the concentration profiles sharpen and the colony height narrows as finer grid resolutions are employed (compare profiles between -0.2 and 0.4 depth). This spreading as the resolution is coarsened allows the acetate utilizing fraction of the colony to grow more quickly and to a larger overall fraction of the total colony composition (see Figure 7.4a). The acetate concentration profile is especially illustrative of the resolution dependence (see Figure 7.3 (right), showing how the structure changes as

the resolution is coarsened. Encouragingly, the profiles do converge as the resolution is increased.

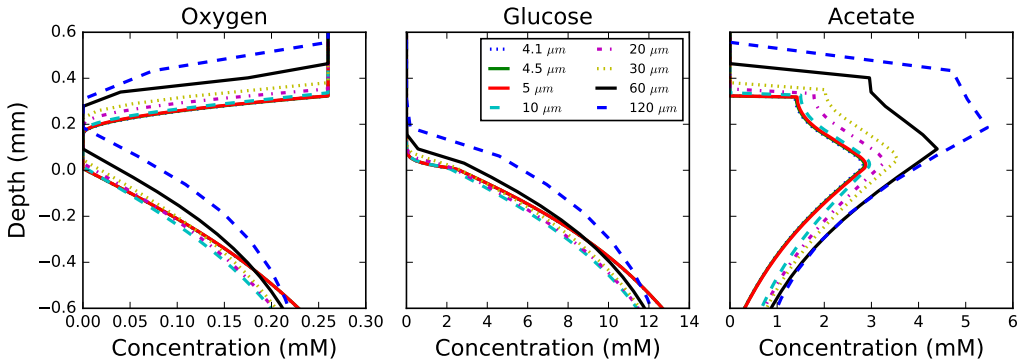


Figure 7.3: **Species Profiles.** Profiles through the center axis of an *E. coli* K-12 MG1655 colony showing the concentrations (x-axis) of several metabolites at a given depth (y-axis) relative to the agar surface after 30 hours of colony growth. Simulations demonstrate that concentrations predicted by 3DdFBA have a strong dependence on the grid resolution. Concentration profiles are essentially converged after 10 µm; the deviation below a depth of about -0.2 mm seen for higher resolutions (4.1 to 5 µm) is due to boundary condition effects as a thinner agar layer was needed to allow for the simulation to fit in GPU memory.

Perhaps more pertinent to questions regarding metabolic cross-feeding is the population composition; therefore, we next examined the error in the predicted fractions of glucose- and acetate-utilizing cells as a function of the grid resolution (Figure 7.4b). As no analytical solution to the problem is known, we compute the error relative to the finest grid resolution simulation (i.e. 4.1 µm). The average relative error was computed over the period from 10 to 30 hours of growth as this is when the initial expansion and differentiation occurs in the colony. The results demonstrate that the error rapidly converges, and that below approximately 10 µm grid spacing, the errors in

the population fractions fall below $< 10\%$ (Figure 7.4b). Of special note is the fact that when resolutions become large, error in the population fractions rapidly increases and in fact some numerical instability—as demonstrated by the non-physical oscillatory solution—emerges with grid resolutions coarser than 20 µm (Figure 7.4a). Previous spatially-resolved FBA studies have examined mixed-species simulations with grid resolutions ranging from 200µm to $\sim$500µm [453]. Our results show that in order to simulate processes in which the reaction to diffusion ratios of metabolites are similar to those of glucose and acetate, finer grid resolutions must be used. While we did not study the resolution dependence of any of the previous studies, our results suggest that at least some of them may not have been simulated at a resolution adequate to ensure reasonably converged results. Nevertheless, we note that our results are for a colony growing at the boundary of two different phases (*i.e.* agar/cells and air), one of which occludes certain metabolites, and that the convergence characteristics might therefore be different from previous studies.

7.3.2 Strain-Dependent Features of Acetate Cross-Feeding

Different *E. coli* strains have evolved different glucose utilization rates (see Table 7.1), presumably due to some environmental (or engineered, in the case of commercial strains) stress. While aerobic and anaerobic growth rates were highly correlated with glucose utilization rates ($p<0.05$ and $p<0.02$, respectively; data not shown), the aerobic and anaerobic growth rates were not significantly correlated ($p>0.19$; data not shown). Guessing what cross-

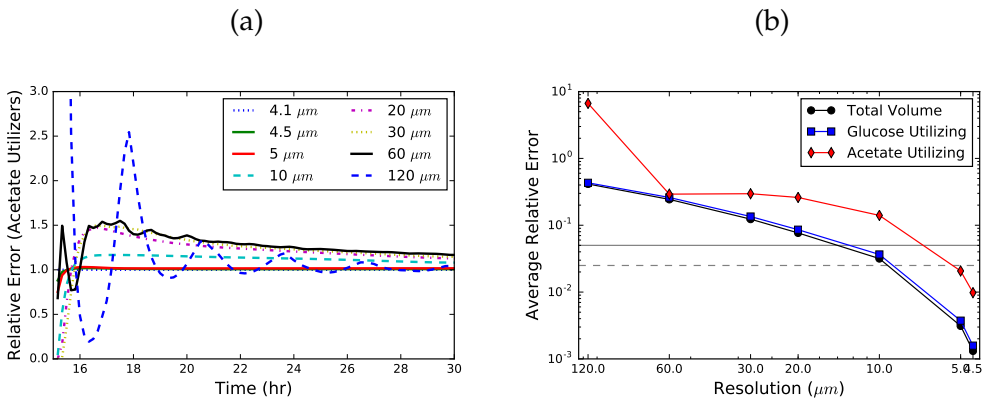


Figure 7.4: **Error Analyses.** a) Error in the acetate utilizing population fraction, relative to the 4.1 µm simulation, after the emergence of metabolic cooperativity (occurring at about 15 hours). Coarse resolutions introduce numerical oscillations when coarser than about 20 µm. b) Error in the computed volume of cells in an *E. coli* K-12 MG1655 colony as a function of grid resolution. Errors are computed as the relative deviation from a simulation with a 4.1 µm grid resolution averaged from 10 to 30 hours of colony growth. Total colony volume (black circles), volume of glucose utilizing cells (blue squares) and volume of acetate utilizing cells (red diamonds) are shown. Horizontal lines show 5% (solid) and 2.5% (dashed) error in the volumes.

feeding behavior will arise due to these differences is nontrivial. To complicate the issue, two of the strains produce acetate via overflow metabolism when grown aerobically, the rates of which are uncorrelated with the acetate production rate when grown anaerobically.

We hypothesized these differences could give rise to strain-specific cross-feeding features which 3DdFBA simulations could predict. Models for each of the organisms were obtained from the BiGG database [458] and fit to measured aerobic and anaerobic growth and acetate secretion rates [457]. The fit models exhibited very low error in the growth and aerobic acetate secretion rates and moderate error in anaerobic acetate production rates (see Table 7.1 and Methods for a detailed analysis of the models). Each strain was

simulated on a flat agar surface with identical initial conditions for 50 hours of growth with a grid resolution of 10 μm to ensure a converged answer.

Overall, colonies grew at different rates due primarily to their differences in substrate utilization rates. Generally, glucose diffusing up through the agar was depleted by the cells growing at the periphery and bottom of the colony, while oxygen diffusing in from above was predominantly consumed by those at the top of the colony (producing an anoxic zone with little growth at the colony center). Heatmaps depicting concentrations and fluxes of the major metabolites in a slice through the colony center can be seen in SI Figure 2. The structure of the actively growing cells (depicted by high metabolic flux in the SI Figure 2) are generally the same, though several significant differences can be seen (for instance, compare oxygen and glucose uptake rates for strains W3110 and BL21). Counter-intuitively, height to width ratios of the colonies varied by at most 22% early on, and settled down to a maximum difference of 12% (see SI Figure 3).

Acetate cross-feeding naturally arose in simulations of all *E. coli* strains (Figure 7.5). While the structural profiles of the colony (*i.e.* the spatial arrangement of acetate to glucose utilizing populations) were similar among the strains, several features did vary significantly. Specifically, the timing of the onset of the cross-feeding and the partitioning of the colony members into metabolic phenotypes differed significantly (see Figure 7.5; top left). The timing of the emergence of an acetate utilizing fraction varied by $\sim$10 hours, ranging from 15 to 25 hours after inoculation. This time appears to primarily be set by the rate of aerobic growth on glucose (*cf.* Table 7.1,

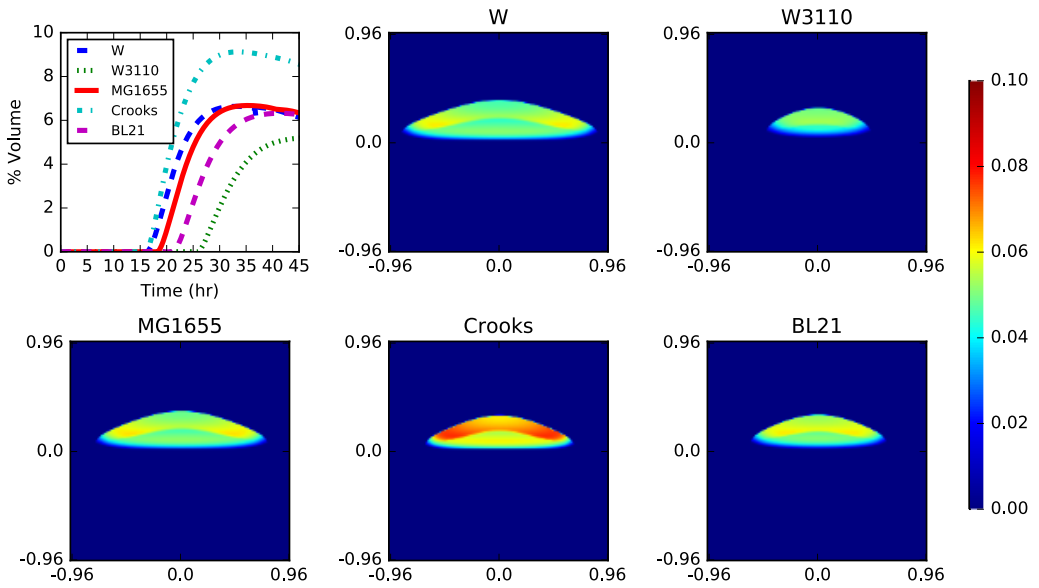


Figure 7.5: **Strain Dependence of Cross-Feeding.** Differences in metabolite utilization efficiency of closely related *E. coli* strains give rise to differences in metabolic cross-feeding. Cross-sections of the acetate utilizing volume fraction of cells after 30 hours of growth show that, while the colony growth rates are slightly different, the structure of the metabolic cross-feeding is essentially the same (heatmaps). A quantification of colony volume, however, shows that the fraction of colony that utilize acetate can vary by a factor of 2 and the timing for onset of metabolic cooperativity can vary by up to 10 hours (top left). The acetate utilizing fraction are cells that have transitioned into the phenotypic state where they can catabolize acetate; while these cells are not necessarily consuming acetate, they have the capacity to do so. In general, about 70% of these cells are consuming acetate and are doing so at the maximum uptake rate.

Figure 7.5 and SI Figure 3). This can be understood simply to be a matter of how much time is required for the colony to grow tall enough such that the glucose entering the bottom is metabolised before it reaches the top. Partitioning among the metabolic cooperators also shows a high degree of strain dependence. After the onset of metabolic cooperativity, the acetate

utilizing fraction of the population quickly rises to some (nearly) steady-state. Acetate utiliziers were found to comprise between 5 and 10% of the colony by volume (of which ∼70% are actually consuming acetate). The Crooks *E. coli* strain had the largest fraction of acetate utilizers while the W3110 strain had the smallest fraction. The other three strains had similar acetate utilizing fractions of about 6 and 7%. Quantitatively, the strains shown in Figure 7.5 have significantly different chemical turnover; for instance Crooks (W3110) has ∼40% higher (lower) acetate turnover after accounting for differences in colony volume (see SI Figure 4).

To identify the sources of differences between strains, correlations were computed between various strain-dependent characteristics and both the maximum acetate utilizing fraction and the onset time (see Table 7.2). We found that the onset time and overall acetate fraction were set by qualitatively different strain features. Specifically, the onset time was primarily controlled by the maximal aerobic growth rate (with faster-growing strains having earlier onset times), while the overall acetate utilizing fraction was only weakly correlated with the aerobic and anaerobic growth rates. The acetate utilizing fraction turned out to be highly anti-correlated with the maximal acetate uptake rate. In essence, the onset time depends on how quickly the colony grows large enough to have an anoxic region in the interior, while the thickness of the acetate utilizing fraction depends on how much of the colony is acetate-rich, which in turn depends on how fast cells deplete the available acetate. We found that gene expression values for malate dehydrogenase (mdh), succinyl-CoA synthettase (sucCD), and 2-oxoglutarate

dehydrogenase (sucAB) were correlated with anaerobic flux values through their reactions (a similar result was identified in [457]), and could be the sources of differences in growth rates and acetate production rates in the different strains. More practically, we found that the overall acetate production rate (i.e. loss of acetate to the environment) was highly correlated with the acetate utilizing fraction.

Subtle differences in strain growth, when considered in the context of microbial communities, could give rise to drastically different population dynamics. For instance, competition for acetate with another microbe could be drastically affected by the onset time for metabolic cross-feeding. Additionally, partitioning of the colony into different phenotypes could effect its robustness to environmental changes or stresses. While we did not study these effects here, it is clear that the 3DdFBA methodology could be used for their investigation.

7.3.3 Geometry Dependence of Acetate Cross-Feeding

The cross-feeding behavior in *E. coli* depends on characteristics intrinsic to the particular strain (*i.e.* growth rate, maximum uptake and efflux rates, metabolic efficiency, *etc.*) as well as interactions with the environment (*i.e.* structural confinement and availability of nutrients). To examine the latter effect, we simulated *E. coli* K-12 MG1655 growing on agar surfaces with various geometric features (see Figure 7.2). A grid resolution of 10 μm were used for all simulations, and the colony was seeded in the center of the computational domain. Characteristic images of simulations of these

geometries can be seen in Figure 7.6. Significant differences in colony growth rate and acetate cross-feeding are apparent in the figure. Specific features of each geometry will be described in turn.

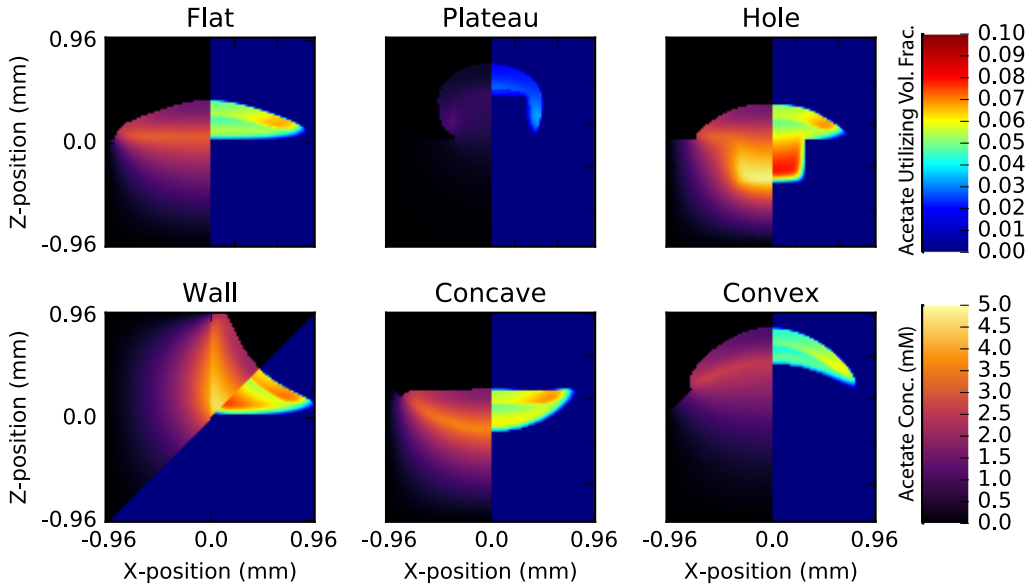


Figure 7.6: **Geometry Dependence of Cross-Feeding.** Snapshots of each of colonies after 35 hours of growth in six different geometries showing the drastic difference in growth rate and acetate cross-feeding caused by agar geometry. Each image shows acetate concentration (left) and volume fraction of acetate utilizing cells (right). The acetate utilizing fraction are cells that have transitioned into the phenotypic state where they can catabolize acetate; while these cellsaare not necessarily consumping acetate, they have the capacity to do so. In general, about between 45 and 76% of these cells are consuming acetate and are doing so at the maximum uptake rate.

Wall

Colonies were grown on an agar surface with starting distances ranging between 0 to 210 µm from a 90 degree agar wall (see Figure 7.7). In this geometry, the colony grows until it interacts with the agar wall, which not

only provides a physical barrier, but also an additional reservoir of glucose to adjacent cells. As a result, colonies seeded closer to the wall tended to grow more quickly in general (and asymmetrically in the direction of the wall), and the rate of colony growth increased after interaction with the wall (up to 10% greater; see Figure 7.7; top left). The onset of acetate cross-feeding varied only by about 3 hours; however, the fraction of total cells in a colony that utilized acetate was relatively unchanged after a about 20 hours of growth (Figure 7.7; bottom left).

Two especially interesting features are seen in the wall geometry simulations. First, while the agar surface area is the same, the colonies tend to grow more quickly than they otherwise would a flat surface. Second, an additional acetate utilizing fraction forms near the wall edge starting after about 20 hours (see Figure 7.6). This acetate fraction grows significantly faster than the one seen in flat surface colonies, and leads to a larger overall acetate volume fraction by about 25%. The formation of this second fraction is due primarily to the large build-up of acetate in the center of the colony. A similar effect is seen in the hole geometry discussed later.

Hole and Plateau

Colonies growing on the top of plateaus and at the bottom of holes exhibit the most complex dynamics of any in this study. For plateaus, the height is the most important role, primarily because glucose becomes severely limited as the column grows taller (*cf.* left vs. right in Figure 7.8). Growth significantly slows after the glucose in the column is consumed (Figure 7.8;

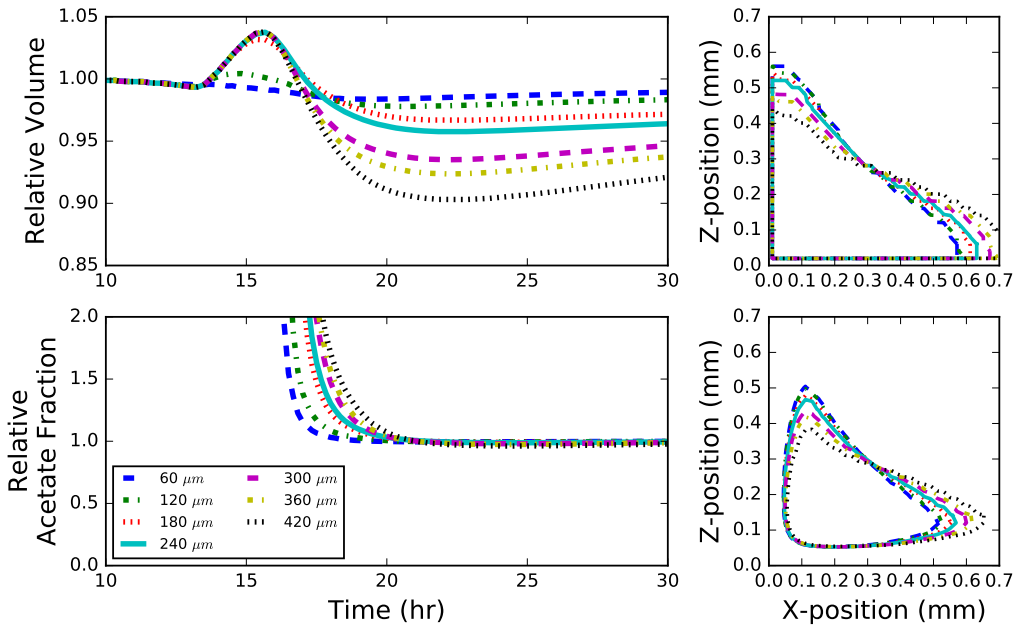


Figure 7.7: **Wall Geometry.** Behavior of *E. coli* colonies grown next to an agar wall containing metabolizable glucose. Different curves indicate the distance of the initial colony seed from the wall. Colonies grown further from the wall generally grow slower than colonies seeded near the wall, however tend to grow more quickly when they interact with the wall as demonstrated by an inflection in each curve (top left). The fraction of acetate utilizers in a community linearly on the distance from the wall (bottom left). A profile the colony showing the acetate utilizing fraction (contour level=0.05; bottom right) and the whole colony (contour level=0.64; top right) after 25 hours of growth. Colony volume and acetate fraction are shown relative to a colony that began growth at the edge where the wall meets the floor.

top right). Shorter columns lead to a significant decreases in the acetate utilizing fractions after the colonies have grown down the sides of the plateau and begin to interact the substrate below. After some time this effect begins to ebb, and the acetate fraction increases again, settling to a steady state near 8% (slightly larger than that of a colony grown on a flat surface). The duration and extent of this transient drop in acetate utilixing fraction depends strongly

on the aspect ratio of the plateau; wider plateaus exhibit more moderate transients (less of a dip, see Figure 7.8; bottom). Wider plateaus also exhibit faster growth rates (Figure 7.8 top), primarily due to the longer expansion time before having to grow around the edge.

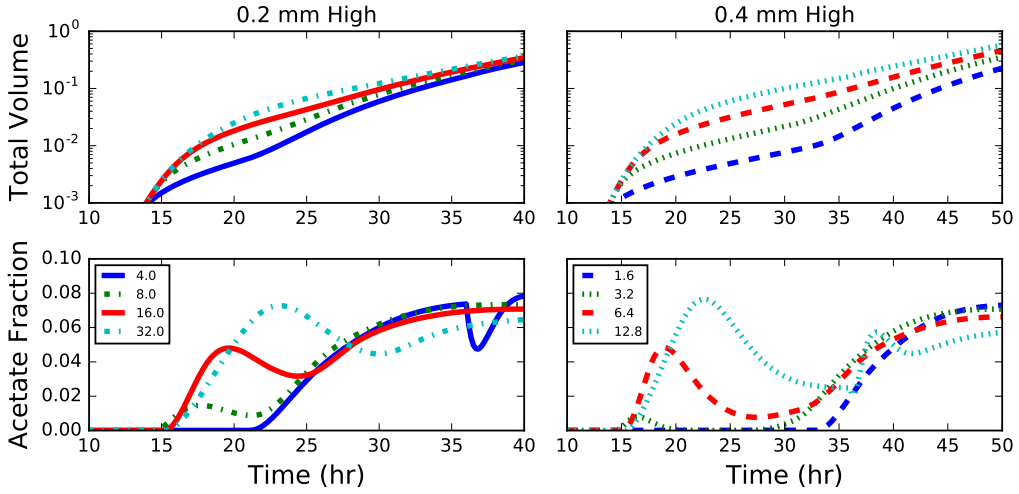


Figure 7.8: **Plateau Geometry.** Plots of total colony volume (top) and fraction of acetate utilizers (bottom) for colonies grown on plateaus of agar 0.2 (left) and 0.4 (right) mm in height. The legend indicates the aspect ratio (width:height) of the plateau and colors correspond to simulations with the same plateau width.

The behavior of colonies that grow in holes, on the other hand, are predicted to depend weakly on hole depth; however the aspect ratio plays a more interesting role. For narrow enough holes (low aspect ratio) the colony interacts with the glucose rich walls before acetate utilizing fractions begin to arise. High aspect ratio holes cause the growth of the acetate utilizing fraction to pause slightly until the colony has crested the edge and the distance from the agar (*e.g.* colony height) again drives growth in the acetate utilizing fraction (see Figure 7.9). The steady-state acetate

utilizing fraction is generally larger than for flat colonies, laying somewhere around 8%. The hole geometry is particularly representative of a wound geometry; nutrients flow from the substrate and the colony top is exposed to oxygen. It has recently been shown in 1D simulations of diseased wound biofilm composed of *Pseudomonas aeruginosa* and *Staphylococcus aureus* that spatial partitioning (and colony stability) is dependent on cross-feeding and nutrient availability [454, 455]. It would be interesting to investigate how a 3D structure more representative of real wounds would change the results of these studies.

In general, holes and plateaus are both predicted to drive larger fractions of the colony into acetate consumption, but these effects are due to different reasons. In holes, a significant fraction of the colony grows very far from any oxygen, allowing acetate to build up inside the colony driving a larger fraction into acetate utilization state. This is quite apparent from examining images of the simulations (see Figure 7.6). When grown on plateaus, the limitation of glucose causes the large acetate utilizing fraction of cells near the column. It is easy to imagine how more complex arrangements of plateaus and holes will exhibit even more complex dynamics. Complex arrangements are easy to implement in microfluidic devices; therefore, testing the results of these simulations—and therefore the validity of the 3DdFBA model—should be relatively straightforward.

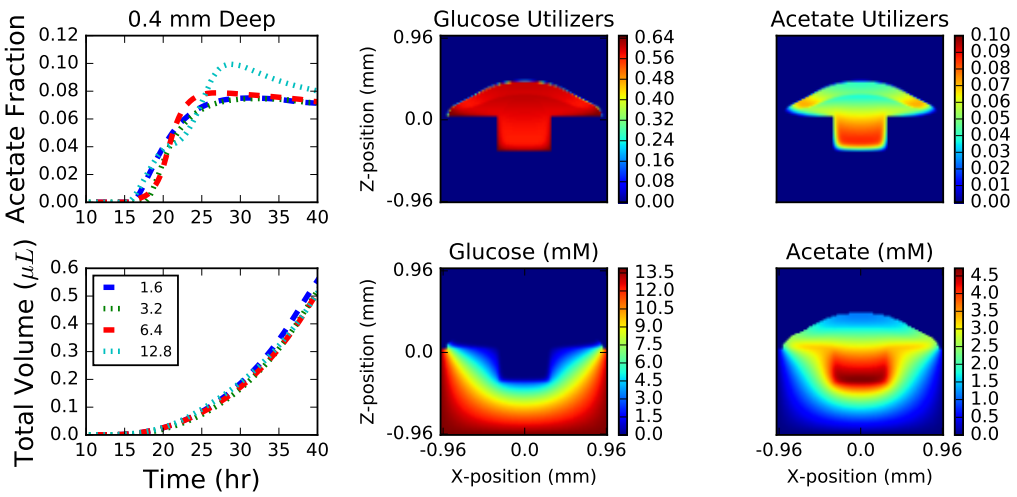


Figure 7.9: **Hole Geometry.** Results for colonies grown at the bottom of a square hole in an agar surface 0.4 mm deep. Results for holes 0.2 mm deep are not shown as they are similar. The colony growth (top left) is mildly dependent upon the width of the hole with wider holes growing slightly more quickly. The relative fraction of acetate (bottom left) utilizers in the population grow with time until the wall the colony reaches the wall and a burst of glucose utilizers are grown until the colony on the wall is thick enough to begin producing acetate utilizers once again. The confinement leads to larger fraction of acetate utilizers than seen on flat surfaces. Snapshots of population volume fractions (top middle and right) and chemical concentrations (top middle and right) after 40 hours of growth. Further, a large build-up in acetate concentration (bottom middle; units of mM) inside the hole is apparent while nothing interesting occurs with the glucose concentration (bottom right; units of mM).

Concave and Convex

Natural surfaces tend to curve. Curvature is especially important in human microbiome research as essentially all internal (*i.e.* gut, mouth, stomach, *etc.*) and external (*i.e.* skin, eye, nose, eardrum *etc.*) surfaces exhibit curvature on the length scales of millimeters to centimeters. We simulated *E. coli* growing on concave and convex agar surfaces with varying curvature. Concave

surfaces are particularly interesting as they are quite analogous to a surface wound; nutrients flow up from inside the body while oxygen flows from above. While *E. coli* is not a major chronic wound pathogen, understanding the effects of curvature are nevertheless interesting. Simulations were run with curvatures ranging from about 0.5 (2mm radius) to about 1.0 (1 mm radius) in a single dimension (cylinder-like geometry).

Colony growth rate was approximately linearly related to curvature over the 30 hours of simulated time (see SI Figures 5 & 6). While higher curvatures introduced relatively small increases in growth on concave surfaces (<2% compared to flat surfaces after 30 hours), they impeded growth on convex surfaces (up to 5% in the same amount of time). Our results demonstrate that a curved surface has relatively insignificant effect on the colony growth when compared to walls, plateaus and holes. This suggests that simulations of biological systems need not necessarily capture the exact geometry of a system, but rather the major features arising abrupt changes to colony confinement.

7.4 Conclusions

Here we have performed the first parametric study of the effects of strain specific features and substrate geometry on the growth of *E. coli*. Using 3DdFBA, a multi-scale method coupling reaction-diffusion with FBA, we were able to elucidate subtle differences in the growth and metabolic cross-feeding of colonies growing on various agar surfaces due to varying initial

conditions. Three observations are particularly important for future studies.

First, by examining the dependence of the solution on grid resolution we determined that a lattice spacing of $\sim$10 µm (or smaller) is required to ensure a converged solution. While larger resolutions get the overall colony growth more or less correct, larger errors in the extent of cross-feeding arise. And in fact, with grid resolutions of greater than about 30 µm, oscillatory patterns arise in the solution, suggesting numerical instability. While we acknowledge that the actual solution depends on the relative rates of diffusion, reaction and regulation, we nevertheless suggest future studies use a grid spacing of $\sim$10 µm when making quantitative predictions about metabolic cross-feeding.

Second, metabolic cross-feeding can depend on strain specific characteristics, even for nearly (genetically) identical organisms. As we found for five *E. coli* strains, the onset time of metabolic cross-feeding and the partitioning of the colony between metabolic phenotypes depend non-linearly on features such as growth, uptake and efflux rates for shared metabolites. These subtle differences could give rise to drastically different behavior when growing in consortia of competing organisms. Hypothetically, if another organism were to compete for acetate, *E. coli* strains W3110 and BL21 might not fractionate into different phenotypes, while the Crooks and W strains might. This behavior will likely be highly dependent on the actual scenario simulated; as a recent study of microbiome associations showed, the extent of cooperation versus competition is highly dependent on the concentration of available nutrients [460]. This conjecture could be easily verified via further 3DdFBA

simulations.

Third, by examining various idealised geometries we were able to identify which features significantly impacted cross-feeding. Colonies growing on "hard" geometries near walls, edges, holes and plateaus resulted in significant differences compared to growth on flat surfaces. Partitioning of the population between phenotypes could be significantly affected by geometry (*e.g.* plateau and hole). Additionally, such geometries could introduce transient deviations from "steady-state" growth which could last for more than 20 hours (depending on the particular geometry). These results are in contrast to those seen for "soft" geometries (convex or concave surfaces), which showed minor deviations of maximally 5% in growth. Overall, our results suggest that when simulating microbial communities, it may suffice to only capture abrupt geometric features (*i.e.* walls, turns, confinement, *etc.*) while neglecting minor features (*i.e.* rough surfaces, mild curvature, *etc.*).

We believe that even this relatively simple study demonstrates the utility of 3DdFBA; such multi-scale methods supplement experimental techniques that are limited in temporal and/or spatial resolution. The simplicity and speed of 3DdFBA (nearly real-time on a modern, inexpensive GPU [58]) means it could become a computational instrument in experimental and theoretical laboratories alike. That being said, there are a number of algorithmic improvements that need to be implemented; such as multi-GPU spatial decompositions (for example [429]), and the addition of an advection equation to the algorithm, such as has been proposed by Chen *et al.* [454].

This work paves the way for future work examining more complex sys-

tems like the human gut microbiome. Recent work already demonstrated the potential for cross-feeding in the gut microbiome in idealized populations [461]. Further, a number of tools/databases that identify metabolic cross-feeding from metagenomic datasets (such as MMinte [460] and AGORA [462]) are now available. These will help inform construction of realistic model gut communities. Such studies could lead to insights into community dynamics, and potentially, their connection to disease.

7.5 Supporting Information

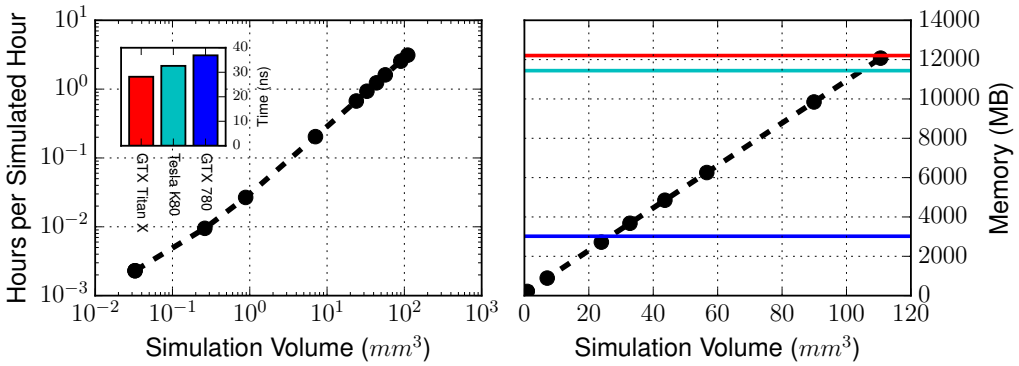


Figure 7.10: **Code Performance.** Benchmarks for the 3DdFBA code used in this manuscript on 3 NVIDIA GPU models: 1) GeForce GTX Titan X, 2) Tesla K80 and 3) GeForce GTX 780. (left) Hours of real time required to simulate an hour of colony growth as a function of the simulation volume. Simulations were performed with a 1 μm lattice spacing. The inset shows the performance of the three different GPUs in units of time (ns) per lattice sites per timestep. (right) Memory required for a simulation consisting of 2 cell types and 5 chemical species. Horizontal lines show memory limitations of the GPUs indicated in the inset. Current GPUs can support simulations of over 100 mm^3.

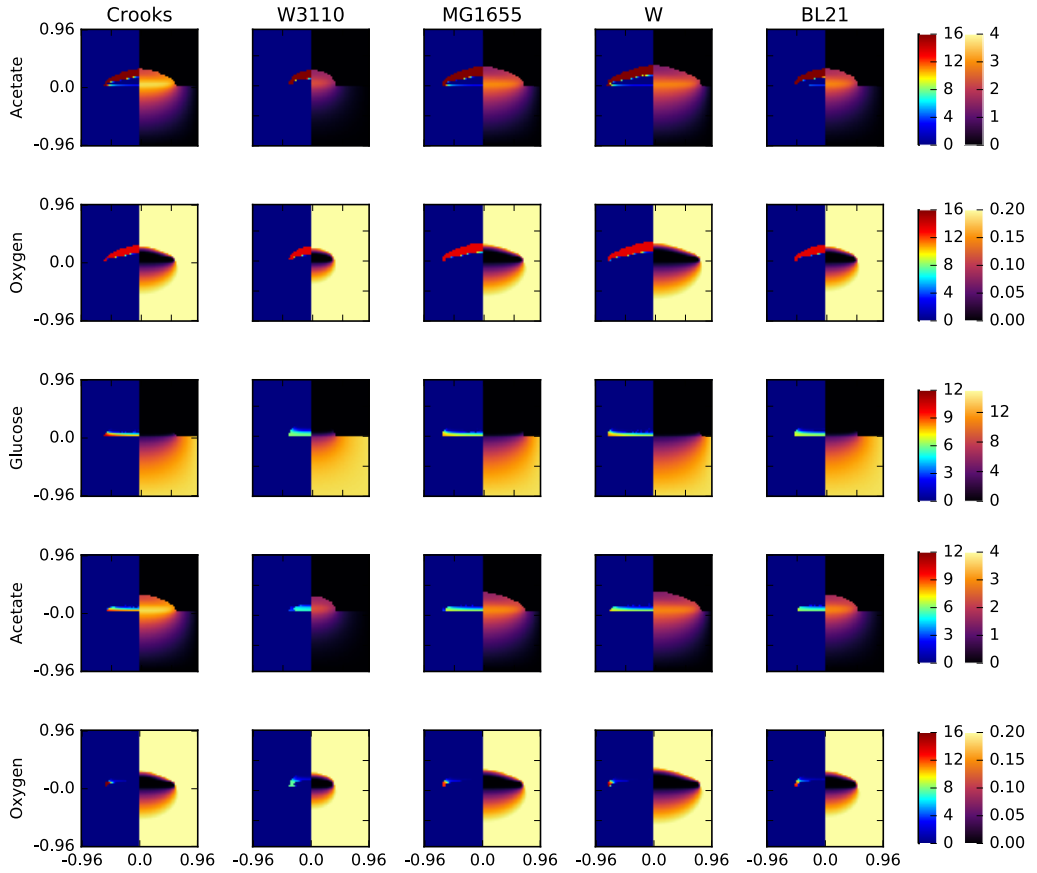


Figure 7.11: **Concentration and Fluxes for Different *E. coli* Strains.** Colony cross-sections for the five strains depicting concentrations (right) and fluxes (left) for key metabolites after 30 hours of growth in acetate utilizing (top two) and glucose utilizing (bottom three) cells. Concentrations are reported in units of mM; fluxes in units of mmol/gDCW/hr.

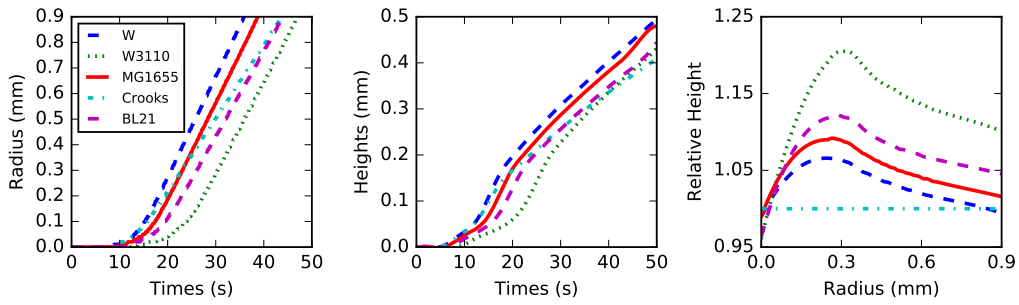


Figure 7.12: **Colony Expansion for Different *E. coli* Strains.** Growth of *E. coli* colony radius (left) and height (middle) initially show an exponential character limited by cell mass, prior to transitioning to a linear regime where nutrient availability is the primary factor limiting growth. Variability in nutrient uptake rates among strains gives rise to differences in the colony aspect ratios when colonies are still relatively small (right).

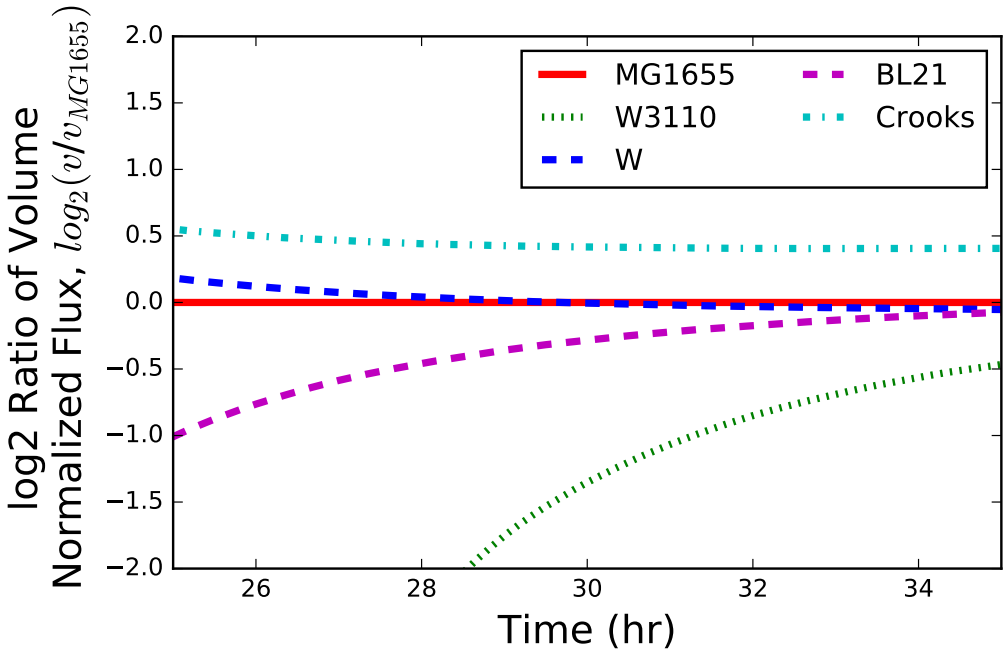


Figure 7.13: **Relative Acetate Turnover.** Colony volume normalized acetate consumption flux (e.g. $mmol/L/hr$) relative to strain MG1655.

Table 7.2: **Growth Correlations to Strain Features.** Correlations of maximal acetate utilizing fraction (left two columns) and onset time of acetate utilization (right two columns) with various bulk strain characteristics. Values that are significant with a (two-tailed) P-value ≤ 0.01 are indicated in bold.

Variable	Pearson r	p-value	Pearson r	p-value
Acetate Production Rate[a]	**0.9912**	**0.0010**	-0.6157	0.2689
Acetate Efflux Rate[b]	**0.9892**	**0.0013**	-0.6083	0.2764
Acetate Influx Rate[c]	**-0.9851**	**0.0022**	0.5941	0.2907
v_{o2} [d]	**-0.9836**	**0.0025**	0.7487	0.1454
$v_{glucose}+\mathbf{O_2}$[d]	-0.9393	0.0178	0.8873	0.0446
$v_{acetate}-\mathbf{O_2}$[d]	0.7805	0.1193	-0.6524	0.2327
$v_{glucose}-\mathbf{O_2}$[d]	-0.7175	0.1724	0.6480	0.2370
$v_{biomass}+\mathbf{O_2}$[d]	0.7043	0.1843	**-0.9884**	**0.0015**
$v_{biomass}-\mathbf{O_2}$[d]	0.4489	0.4482	-0.5885	0.2966

[a]Total rate of acetate loss to the environment by the whole colony (after 35 hours of growth). [b]Rate of acetate efflux integrated over the whole colony (after 35 hours of growth). [c]Rate of acetate uptake integrated over the whole colony (after 35 hours of growth). [d]Correlation was computed with the maximal possible uptake/secretion flux for aerobic ($+\mathbf{O_2}$) and anaerobic growth ($-\mathbf{O_2}$).

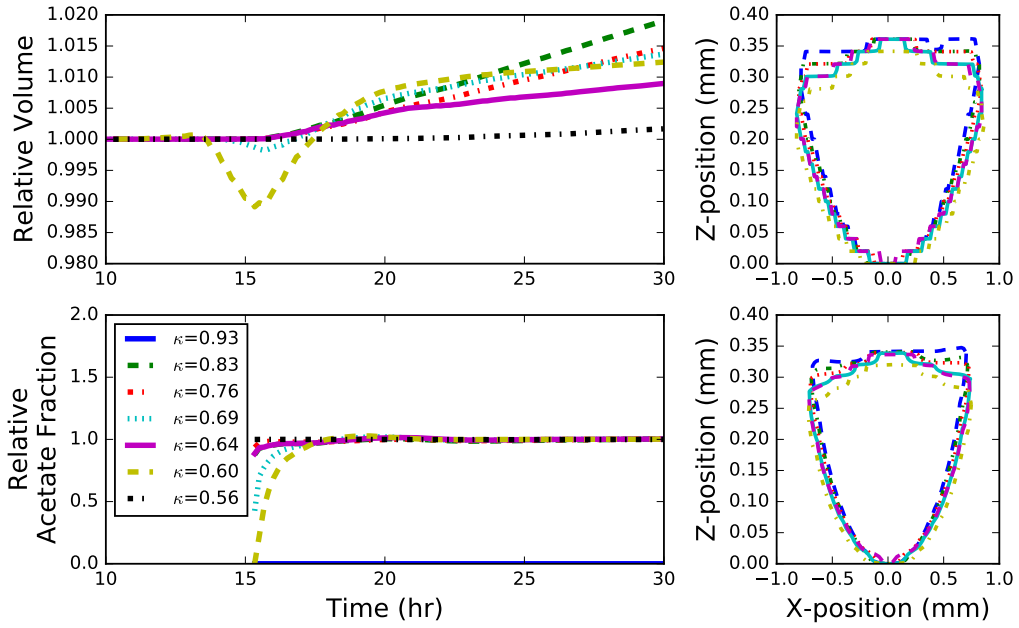


Figure 7.14: **Concave Surface.** Behavior of *E. coli* colonies grown on a concave surface containing metabolizable glucose. Different curves indicate different curvatures of the substrate. (top left) Colonies grown on substrate with higher curvature grew slightly faster after many hours of growth. (top left) The fraction of acetate utilizers in a community (relative to a colony growing on a nearly flat surface) depends linearly, but inconsequentially on the substrate curvature. A profile the colony showing the acetate utilizing fraction (contour level=0.05; bottom right) and the whole colony (contour level=0.64; top right) after 35 hours of growth. Colony volume and acetate fraction are shown relative to a colony grown on a surface with a curvature $\kappa = 1.92$.

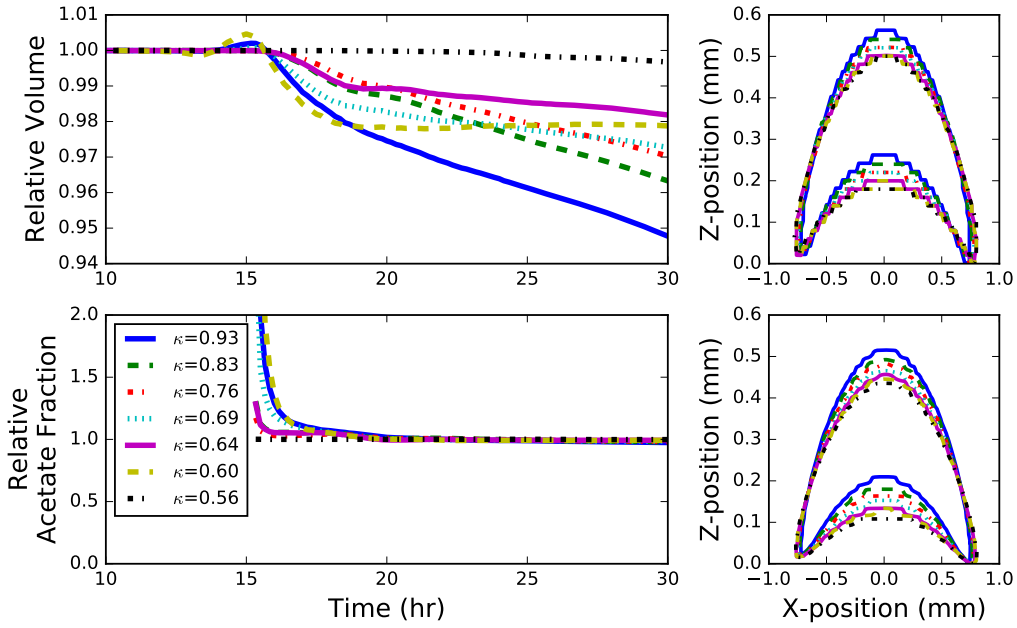


Figure 7.15: **Convex Surface.** Behavior of *E. coli* colonies grown on a convex surface containing metabolizable glucose. Different curves indicate different curvatures of the substrate. (top left) Colonies grown on substrate with higher curvature grew slightly slower after many hours of growth. (top left) The fraction of acetate utilizers in a community (relative to a colony growing on nearly flat surface) depends linearly, but inconsequentially on the substrate curvature. A profile the colony showing the acetate utilizing fraction (contour level=0.05; bottom right) and the whole colony (contour level=0.64; top right) after 35 hours of growth. Colony volume and acetate fraction are shown relative to a colony grown on a surface with a curvature $\kappa = 1.92$.

Chapter 8

Conclusions

The investigations presented in this thesis are six vignettes demonstrating on how computation applied to biology can distil essential features from the complexity and provide hypotheses that could guide targeted experimentation. Different methodologies were leveraged in each study as the phenomena studied varied in temporal and spatial scale. Together, they demonstrate how a scientist can tailor the approach to the problem posed, rather than finding the nail to hit with their favourite method. I will briefly summarize the work and conclude with my estimation of the state and direction of the field.

8.1 Summary

In Part I of the thesis, studies of *Methanosarcina* species were presented. A kinetic model of methanogenesis—the metabolic pathways unique to Archaea that produce methane—in *M. acetivorans* was developed and used to examine the sensitivity of methane production rates to abundances of methanogenesis proteins. Subsequently, the metabolism of *M. acetivorans* when grown on several substrates was examined using genome-scale metabolic modeling. Metabolic phenotypes, wherein the methanogens utilize metabolic pathways to different extents, were predicted by integrating RNA expression and half-

life data with the models. Strikingly, it was shown that the organism adjusts RNA half-lives of nearly half of metabolic genes to optimize metabolic flux for different growth substrates. This discovery was the first to show such a global role for half-life in defining metabolic phenotype. Concomitantly, the metabolic model was corrected and expanded, especially in the context of the cell's compositional requirements, by adding new terms to the model's biomass equation. Two comparative genomic studies were subsequently undertaken, enabled primarily by the availability of this and other highly curated metabolic models. First, the genomes of all fully sequenced Archaea were mapped across the available metabolic models to examine conservation of metabolic function. This revealed that amino acid metabolic pathways relatively more highly conserved than coenzyme, lipid, nitrogen, and transport metabolism. Second, the metabolic models of several *Methanosarcina* species were mapped across the genomes of 30 *Methanosarcina* species, enabling a pan-reactome study of these metabolically diverse methanogens. By examining the resulting core-reactome in the context the conserved genome, knowledge gaps in the metabolism could be filled. Importantly, by examining the pan- and core-genome of the *Methanosarcina*, a biosynthetic pathway for methanophenazine, a methanogenesis cofactor, was hypothesized.

In Part II of the thesis, causes of stochasticity and heterogeneity were examined in the model organism *E. coli*. Adopting a simulation technique designed to sample the chemical master equation noise in gene expression was examined. Inspired by the inability of traditional models (which neglected genome replication) to fit the distributions obtained using single-molecular

fluorescence in situ hybridization experiments, the effect of genome replication on the noise observed in genes placed at different locations around the circular genome was examined. Simulation results indicated that relaxation of the RNA count from a pre- to a post-replication steady-state significantly affected the shape of the resulting distributions. Analytical results showed that the noise of a constitutively expressed gene could be completely defined by three variables: the location of the gene on the chromosome, the RNA half-life, and the cell doubling time. Overall, this showed that previous studies that neglected to handle genome replication explicitly could both qualitatively and quantitatively misinterpret experimental data. Finally, adopting a completely different modeling framework, metabolic cooperativity in *E. coli* colonies growing on agar surfaces were examined. Building upon previous work that identified acetate cross-feeding using reaction-diffusion partial-differential equations, the effects of strain specific differences in the metabolic capacities and geometrical confinement were examined. The behavior of five different *E. coli* strains were examined; it was found that the extent and timing of metabolic cross-feeding were significantly different, even for closely related strains. Finally, cross-feeding was found to only vary when the growth substrate had abrupt changes in geometry (e.g. a wall or pit), and that smooth changes caused imperceptible changes in growth.

8.2 Outlook

There is no doubt that the field of computational biology will continue to grow. Increasing computer and new algorithms/methods will allow lager and more sophisticated investigations into biological processes. While some fields have gained traction, by in large computational biology at the level of cells and colonies is not taken seriously by many biologists. There are four hurdles that need to be overcome before computational biology starts to be taken seriously: 1) problems need to be biology-driven rather than model-driven, 2) models need to be predictive rather than reflexive, 3) simulation needs to drive experiment and be performed in an iterative fashion, and 4) model validation/verification and uncertainty quantification will need to become commonplace. I will explain what I mean by these in turn.

First, it should be self-evident that the biological question should drive the biological modeling. Unfortunately, too often models are constructed for modeling-sake. This is often to push the limits of current approaches or to merely document a biological phenomena in a mathematical form. In principle, there is nothing wrong with this, but if modeling does not provide answers to questions that people care about, it will be hard convince biologists that it is worth doing. The community, myself included, needs to do a better job of sitting down with biologists to explore what outstanding questions they cannot answer due to experimental limitations that will ultimately help propel their research forward.

Points 2 & 3 are special cases of the first hurdle. Models need to be able

to make predictions. Often a model is produced and parametrized based some experimental data, and then the fact that the model recapitulates the experimental data is touted. I do not intend to downplay the difficulty in model production and fitting; these are challenging tasks. At the end of the day, the model needs to predict a phenomena that is not facilely obtainable with an experiment to justify the effort. This leads to the third point, that modeling should iterate with experiment: predictions should be verified or refuted and the results reincorporated into the model to push predictive capability further. This iterative process will have the side effect of bring the experimentalist and constitutionalist closer together such that they are better posed to make important discoveries.

Fourth, the models need to be taken seriously. I read this recently in a paper [463], and loved the premise. One of the most critical points in taking a model seriously is verification/validation and uncertainty quantification. Most researchers are good about verification and validation; the former usually means making sure the model recapitulates the experiments on which it is based, while the latter is self-evident—if the model implementation is incorrect it will give the wrong answers. Uncertainty quantification (UQ) is a much trickier, and less common process. UQ attempts to assign error bars or confidence intervals to model predictions or ascribe uncertainties to model inputs. This process is more common in engineering disciplines and computational scientists could learn from this. Taking the predictions of the models seriously and conveying the confidence in those predictions will be essential to being taken seriously by biologist collaborators.

The reason I write these critiques is that I have been guilty of each and every one. Indeed, much of the model building in this thesis falls into one or the other of these categories, with a few predictions here and there. My personal goal for the future is to hold these four considerations in mind when I perform computation. Where I feel the work is strongest is in the breadth of approaches applied to a diverse set of problems, and for this I am proud. One of the key steps forward will be to integrate the various modeling approaches to provide confident and predictive analyses of how biological complexity gives rise to the phenomena of life.

References

[1] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, Nov 1977.

[2] Anja Spang and Thijs J. G. Ettema. Archaeal evolution: The methanogenic roots of archaea. *Nature Microbiology*, 2(8):17109, jul 2017.

[3] W. J. Jones, J. A. Leigh, F. Mayer, C. R. Woese, and R. S. Wolfe. Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Archives of Microbiology*, 136(4):254–261, dec 1983.

[4] D. S. Nichols, M. R. Miller, N. W. Davies, A. Goodchild, M. Raftery, and R. Cavicchioli. Cold adaptation in the antarctic archaeon methanococcoides burtonii involves membrane lipid unsaturation. *Journal of Bacteriology*, 186(24):8508–8515, dec 2004.

[5] R.K. Thauer, A.-Kristin Kaster, H. Seedorf, W. Buckel, and R. Hedderich. Methanogenic archaea: ecologically relevant differences in energy conservation. *Nature Reviews Microbiology*, 6(8):579–591, Jun 2008.

[6] Nadia Gaci, Guillaume Borrel, William Tottey, Paul William O'Toole, and Jean-François Brugère. Archaea and the human gut: New beginning of an old story. *World Journal of Gastroenterology*, 20(43):16062, 2014.

[7] Kristina Lang, Jörg Schuldes, Andreas Klingl, Anja Poehlein, Rolf Daniel, and Andreas Brune. New mode of energy metabolism in the seventh order of methanogens as revealed by comparative genome analysis of "candidatus methanoplasma termitum". *Applied and Environmental Microbiology*, 81(4):1338–1352, dec 2014.

[8] David P. Chynoweth. Environmental impact of biomethanogenesis. *Environmental Monitoring and Assessment*, 42(1-2):3–18, sep 1996.

[9] D.E. Holmes and J.A. Smith. Biologically produced methane as a renewable energy source. In *Advances in Applied Microbiology*, pages 1–61. Elsevier, 2016.

[10] Meisam Tabatabaei, Raha Abdul Rahim, Norhani Abdullah, André-Denis G. Wright, Yoshihito Shirai, Kenji Sakai, Alawi Sulaiman, and Mohd Ali Hassan. Importance of the methanogenic archaea populations in anaerobic wastewater treatments. *Process Biochemistry*, 45(8):1214–1225, Aug 2010.

[11] É. Bapteste, C. Brochier, and Y. Boucher. Higher-level classification of the archea: evolution of methanogenesis and methanogens. *Archeae*, 1(5):353–363, 2005.

[12] Iain Anderson, Luke E. Ulrich, Boguslaw Lupa, Dwi Susanti, Iris Porat, Sean D. Hooper, Athanasios Lykidis, Magdalena Sieprawska-Lupa, Lakshmi Dharmarajan, Eugene Goltsman, Alla Lapidus, Elizabeth Saunders, Cliff Han, Miriam Land, Susan Lucas, Biswarup Mukhopadhyay, William B. Whitman, Carl Woese, James Bristow, and Nikos Kyrpides. Genomic characterization of methanomicrobiales reveals three classes of methanogens. *PLoS ONE*, 4(6):e5797, jun 2009.

[13] P. V. Welander and W. W. Metcalf. Loss of the mtr operon in methanosarcina blocks growth on methanol, but not methanogenesis, and reveals an unknown methanogenic pathway. *Proceedings of the National Academy of Sciences*, 102(30):10664–10669, jul 2005.

[14] J. E. Galagan, C. Nusbaum, A. Roy, M. G. Endrizzi, P. Macdonald, W. FitzHugh, S. Calvo, R. Engels, S. Smirnov, D. Atnoor, A. Brown, N. Allen, J. Naylor, N. Stange-Thomann, K. DeArellano, R. Johnson, L. Linton, P. McEwan, K. McKernan, J. Talamas, A. Tirrell, W. Ye, A. Zimmer, R. D. Barber, I. Cann, D. E. Graham, D. A. Grahame, A. M. Guss, R. Hedderich, C. Ingram-Smith, H. C. Kuettner, J. A. Krzycki, J. A. Leigh, W. Li, J. Liu, B. Mukhopadhyay, J. N. Reeve, K. Smith, T. A. Springer, L. A. Umayam, O. White, R. H. White, E. Conway de Macario, J. G. Ferry, K. F. Jarrell, H. Jing, A. J. Macario, I. Paulsen, M. Pritchett, K. R. Sowers, R. V. Swanson, S. H. Zinder, E. Lander, W. W. Metcalf, and B. Birren. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome Research*, 12(4):532–542, Apr 2002.

[15] M. Rother and W.W. Metcalf. Anaerobic growth of *Methanosarcina acetivorans* C2A on carbon monoxide: an unusual way of life for a methanogenic archaeaon. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16929–16934, 2004.

[16] A. Bose, M.A. Pritchett, M. Rother, and W.W. Metcalf. Differential Regulation of the Three Methanol Methyltransferase Isozymes in *Methanosarcina acetivorans* C2A. *Journal of Bacteriology*, 188(20):7274–7283, 2006.

[17] A. Bose and W.W. Metcalf. Distinct regulators control the expression of methanol methyltransferase isozymes in *Methanosarcina acetivorans* C2A. *Molecular Microbiology*, 67(3):649–661, 2008.

[18] A. Bose, G. Kulkarni, and W.W. Metcalf. Regulation of putative methylsulphide methyltransferases in *Methanosarcina acetivorans* C2A. *Molecular Microbiology*, 74(1):227–238, Oct 2009.

[19] A. Bose, M. A. Pritchett, and W. W. Metcalf. Genetic analysis of the methanol- and methylamine-specific methyltransferase 2 genes of *methanosarcina acetivorans* c2a. *Journal of Bacteriology*, 190(11):4017–4026, 2008.

[20] G. Kulkarni, D. M. Kridelbaugh, A. M. Guss, and W. W. Metcalf. Hydrogen is a preferred intermediate in the energy-conserving electron transport chain of methanosarcina barkeri. *Proceedings of the National Academy of Sciences*, 106(37):15915–15920, sep 2009.

[21] L. Rohlin and R. Gunsalus. Carbon-dependent control of electron transfer and central carbon pathway genes for methane biosynthesis in the Archaean, *Methanosarcina acetivorans* strain C2A. *BMC Microbiology*, 10(1):62, 2010.

[22] N.R. Buan and W.W. Metcalf. Methanogenesis by *Methanosarcina acetivorans* involves two structurally and functionally distinct classes of heterodisulfide reductase. *Molecular Microbiology*, 74(4):843–853, 2010.

[23] N. Matschiavelli, E. Oelgeschlager, B. Cocchiararo, J. Finke, and M. Rother. Function and Regulation of Isoforms of Carbon Monoxide Dehydrogenase/Acetyl Coenzyme A Synthase in *Methanosarcina acetivorans*. *Journal of Bacteriology*, 194(19):5377–5387, Aug 2012.

[24] Nicole Matschiavelli and Michael Rother. Role of a putative tungsten-dependent formylmethanofuran dehydrogenase in *Methanosarcina acetivorans*. *Archives of Microbiology*, 197(3):379–388, Dec 2014.

[25] H. Fu and W.W. Metcalf. Genetic Basis for Metabolism of Methylated Sulfur Compounds in Methanosarcina Species. *Journal of Bacteriology*, 197(8):1515–1524, Feb 2015.

[26] M. J. Reichlen, K. S. Murakami, and J. G. Ferry. Functional Analysis of the Three TATA Binding Protein Homologs in *Methanosarcina acetivorans*. *Journal of Bacteriology*, 192(6):1511–1517, Jan 2010.

[27] M. J. Reichlen, V. R. Vepachedu, K. S. Murakami, and J. G. Ferry. MreA Functions in the Global Regulation of Methanogenic Pathways in *Methanosarcina acetivorans*. *mBio*, 3(4):e00189–12–e00189–12, Jun 2012.

[28] W. W. Metcalf, J. K. Zhang, X. Shi, and R. S. Wolfe. Molecular, genetic, and biochemical characterization of the serC gene of *Methanosarcina barkeri* Fusaro. *Journal of Bacteriology*, 178(19):5797–5802, Oct 1996.

[29] Adam M. Guss, Michael Rother, Jun Kai Zhang, Gargi Kulkkarni, and William W. Metcalf. New methods for tightly regulated gene expression and highly efficient chromosomal integration of cloned genes for methanosarcina species. *Archaea*, 2(3):193–203, 2008.

[30] Adam M Feist, Johannes C M Scholten, Bernhard Ø Palsson, Fred J Brockman, and Trey Ideker. Modeling methanogenesis with a genome-scale metabolic reconstruction of methanosarcina barkeri. *Molecular Systems Biology*, 2, Jan 2006.

[31] Vinay Satish Kumar, James G Ferry, and Costas D Maranas. Metabolic reconstruction of the archaeon methanogen methanosarcina acetivorans. *BMC Systems Biology*, 5(1):28, 2011.

[32] M. N. Benedict, M. C. Gonnerman, W. W. Metcalf, and N. D. Price. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon *Methanosarcina acetivorans* C2A. *Journal of Bacteriology*, 194(4):855–865, Dec 2012.

[33] Q. Li, L. Li, T. Rejtar, B.L. Karger, and J.G. Ferry. Proteome of *Methanosarcina acetivorans* Part II: Comparison of Protein Levels in Acetate- and Methanol-Grown Cells. *Journal of Proteome Research*, 4(1):129–135, 2005.

[34] L. Li, Q. Li, L. Rohlin, U. Kim, K. Salmon, T. Rejtar, R.P. Gunsalus, B.L. Karger, and J.G. Ferry. Quantitative Proteomic and Microarray Analysis of the Archaeon *Methanosarcina acetivorans* Grown with Acetate versus Methanol. *Journal of Proteome Research*, 6(2):759–771, 2007.

[35] Joseph R. Peterson, Piyush Labhsetwar, Jeremy R. Ellermeier, Petra R. A. Kohler, Ankur Jain, Taekjip Ha, William W. Metcalf, and Zaida Luthey-Schulten. Towards a computational model of a methane producing archaeum. *Archaea*, 2014:1–18, 2014.

[36] Joseph R. Peterson, ShengShee Thor, Lars Kohler, Petra R.A. Kohler, William W. Metcalf, and Zaida Luthey-Schulten. Genome-wide gene expression and rna half-life measurements allow predictions of regulation and metabolic behavior in *methanosarcina acetivorans*. *BMC Genomics*, 17(1):924, Nov 2016.

[37] ShengShee Thor, Joseph R. Peterson, and Zaida Luthey-Schulten. Genome-scale metabolic modeling of archaea lends insight into diversity of metabolic function. *Archaea*, 2017:1–18, 2017.

[38] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800, sep 2002.

[39] Peter S. Swain. Efficient Attenuation of Stochasticity in Gene Expression Through Post-transcriptional Control. *Journal of Molecular Biology*, 344(4):965–976, dec 2004.

[40] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, aug 2002.

[41] Nir Friedman, Long Cai, and X. Sunney Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16), oct 2006.

[42] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS Biology*, 4(10):e309, 2006.

[43] Elijah Roberts, Andrew Magis, Julio O. Ortiz, Wolfgang Baumeister, and Zaida Luthey-Schulten. Noise Contributions in an Inducible Genetic Switch: A Whole-Cell Simulation Study. *PLoS Computational Biology*, 7(3):e1002010, mar 2011.

[44] Michael Assaf, Elijah Roberts, and Zaida Luthey-Schulten. Determining the Stability of Genetic Switches: Explicitly Accounting for mRNA Noise. *Physical Review Letters*, 106(24), jun 2011.

[45] Tyler M Earnest, Elijah Roberts, Michael Assaf, Karin Dahmen, and Zaida Luthey-Schulten. DNA looping increases the range of bistability in a stochastic model of the lac genetic switch. *Physical Biology*, 10(2):026002, feb 2013.

[46] Michael Assaf, Elijah Roberts, Zaida Luthey-Schulten, and Nigel Goldenfeld. Extrinsic Noise Driven Phenotype Switching in a Self-Regulating Gene. *Physical Review Letters*, 111(5), jul 2013.

[47] Christopher V. Rao, Denise M. Wolf, and Adam P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237, nov 2002.

[48] Bhaswar Ghosh, Rajesh Karmakar, and Indrani Bose. Noise characteristics of feed forward loops. *Physical Biology*, 2(1):36–45, mar 2005.

[49] D. L. Jones, R. C. Brewster, and R. Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536, dec 2014.

[50] Joseph R. Peterson, John A. Cole, Jingyi Fei, Taekjip Ha, and Zaida A. Luthey-Schulten. Effects of dna replication on mrna noise. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52):15886–15891, 2015.

[51] John A. Cole and Zaida Luthey-Schulten. Careful accounting of extrinsic noise in protein expression reveals correlations among its sources. *Physical Review E*, 95(6), jun 2017.

[52] David Fange and Johan Elf. Noise-induced min phenotypes in e. coli. *PLoS Computational Biology*, 2(6):e80, 2006.

[53] M. Earnest, Tyler, Jonathan Lai, Ke Chen, J. Hallock, Michael, R. Williamson, James, and Zaida Luthey-Schulten. Towards a whole-cell model of ribosome biogenesis: Kinetic modeling of ssu assembly. *Biophysical Journal*, 2015.

[54] Tyler M. Earnest, John A. Cole, Joseph R. Peterson, Michael J. Hallock, Thomas E. Kuhlman, and Zaida Luthey-Schulten. Ribosome biogenesis

in replicating cells: Integration of experiment and theory. *Biopolymers*, 105(10):735–751, jul 2016.

[55] Tyler M. Earnest, Reika Watanabe, John E. Stone, Julia Mahamid, Wolfgang Baumeister, Elizabeth Villa, and Zaida Luthey-Schulten. Challenges of integrating stochastic dynamics and cryo-electron tomograms in whole-cell simulations. *The Journal of Physical Chemistry B*, 121(15):3871–3881, mar 2017.

[56] Piyush Labhsetwar, John Andrew Cole, Elijah Roberts, Nathan D. Price, and Zaida A. Luthey-Schulten. Heterogeneity in protein expression induces metabolic variability in a modeledEscherichia colipopulation. *Proceedings of the National Academy of Sciences of the United States of America*, 110(34):14006–14011, aug 2013.

[57] Piyush Labhsetwar, Marcelo C. R. Melo, John A. Cole, and Zaida Luthey-Schulten. Population FBA predicts metabolic phenotypes in yeast. *PLOS Computational Biology*, 13(9):e1005728, sep 2017.

[58] John A. Cole, Lars Kohler, Jamila Hedhli, and Zaida Luthey-Schulten. Spatially-resolved metabolic cooperativity within dense bacterial colonies. *BMC Systems Biology*, 9(1):15, 2015.

[59] Joseph R. Peterson, John A. Cole, and Zaida Luthey-Schulten. Parametric studies of metabolic cooperativity in escherichia coli colonies: Strain and geometric confinement effects. *PLoS ONE*, 12(8):1–19, 08 2017.

[60] A. Kolmogorov, I. Petrovsky, and N. Piscounov. Étude de l'equation de la diffusion avec croissance de la quantité de matiere et son application a un probleme biologique. *Moscow Univ. Math. Bull.*, 1, 1937.

[61] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, Jan 2010.

[62] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry (North-Holland Personal Library)*. North Holland, 2011.

[63] Volker Grimm and Steven F. Railsback. *Individual-based Modeling and Ecology (Princeton Series in Theoretical and Computational Biology)*. Princeton University Press, 2005.

[64] Ali Khodayari, Ali R. Zomorrodi, James C. Liao, and Costas D. Maranas. A kinetic model of escherichia coli core metabolism satisfying multiple sets of mutant flux data. *Metabolic Engineering*, 25:50–62, sep 2014.

[65] Ali Khodayari and Costas D. Maranas. A genome-scale escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communications*, 7:13806, dec 2016.

[66] Peter D. Karp and Monica Riley. EcoCyc: The resource and the lessons learned. In *Bioinformatics: Databases and Systems*, pages 47–62. Kluwer Academic Publishers, 1999.

[67] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000.

[68] R. Overbeek. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17):5691–5702, sep 2005.

[69] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, mar 2010.

[70] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. Konig, R. D. Smith, and B. O. Palsson. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.*, 6:390, Jul 2010.

[71] Caroline Colijn, Aaron Brandes, Jeremy Zucker, Desmond S. Lun, Brian Weiner, Maha R. Farhat, Tan-Yun Cheng, D. Branch Moody, Megan Murray, and James E. Galagan. Interpreting expression data with metabolic flux models: Predicting mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, 5(8):e1000489, aug 2009.

[72] Sriram Chandrasekaran and Nathan D. Price. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks inEscherichia coliandMycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 107(41):17845–17850, sep 2010.

[73] Anna S. Blazier and Jason A. Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*, 3, 2012.

[74] Brian J. Schmidt, Ali Ebrahim, Thomas O. Metz, Joshua N. Adkins, Bernhard Ø. Palsson, and Daniel R. Hyduke. GIM3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, 29(22):2900–2908, aug 2013.

[75] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, 2006.

[76] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, dec 1977.

[77] Elijah Roberts, John E. Stone, and Zaida Luthey-Schulten. Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. *Journal of Computational Chemistry*, 34(3):245–255, sep 2013.

[78] Joseph R Peterson, Michael J Hallock, John A Cole, and Zaida Luthey-Schulten. A problem solving environment for stochastic biological simulations. *Proceedings High Performance Computing Networking, Storage and Analysis Companion (SCC)*, 2013.

[79] Radhakrishnan Mahadevan, Jeremy S. Edwards, and Francis J. Doyle. Dynamic flux balance analysis of diauxic growth in escherichia coli. *Biophysical Journal*, 83(3):1331–1340, sep 2002.

[80] George Boole. *A Treatise On The Calculus of Finite Differences*. Macmillan and Company, London, 2 edition, 1872.

[81] Julian R. Lebenhaft and Raymond Kapral. Diffusion-controlled processes among partially absorbing stationary sinks. *Journal of Statistical Physics*, 20(1):25–56, jan 1979.

[82] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4576–4579, Jun 1990.

[83] S Winker and C R Woese. A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. *Systematic and Applied Microbiology*, 14(4):305–310, 1991.

[84] G.E. Fox, L.J. Magrum, W.E. Balch, R.S. Wolfe, and C.R. Woese. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proceedings of the National Academy of Sciences of the United States of America*, 74(10):4537–4541, 1977.

[85] C R Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–271, Jun 1987.

[86] R R Gutell and C R Woese. Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs. *Proceedings of the National Academy of Sciences of the United States of America*, 87(2):663–667, 1990.

[87] E. Roberts, A. Sethi, J. Montoya, C. R. Woese, and Z. Luthey-Schulten. Molecular signatures of ribosomal evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37):13953–13958, Sep 2008.

[88] C Woese. The universal ancestor. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6854–6859, 1998.

[89] Carl R Woese. On the evolution of cells. *Proceedings of the National Academy of Sciences of the United States of America*, 99(13):8742–8747, 2002.

[90] C R Woese, G J Olsen, M Ibba, and D Soll. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and Molecular Biology Reviews*, 64(1):202–236, 2000.

[91] P. O'Donoghue and Z. Luthey-Schulten. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiology and Molecular Biology Reviews*, 67(4):550–573, dec 2003.

[92] P. O'Donoghue, A. Sethi, C. R. Woese, and Z. A. Luthey-Schulten. The evolutionary history of Cys-tRNA^{Cys} formation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):19003–19008, Dec 2005.

[93] J.N. Reeve, J. Nölling, R.M. Morgana, and R.R. Smith. Methanogenesis: genes, genomes, and who's on first? *Journal of Bacteriology*, 179(19):5975–5986, 1997.

[94] J.R. Brown and W.F. Doolittle. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and Molecular Biology Reviews*, 61(4):456–502, 1997.

[95] D Graham, R Overbeek, G Olsen, and C Woese. An archaeal genomic signature. *Proceedings of the National Academy of Sciences of the United States of America*, 97(7):3304–3308, 2000.

[96] B. Gao and R.S. Gupta. Phylogenomic analysis of proteins that are distinctive of *Archaea* and its main subgroups and the origin of methanogenesis. *BMC Genomics*, 8:86, 2007.

[97] F. Gailard, B. Scaillet, and N.T. Arndt. Atmospheric oxygenation caused by a change in volcanic degassing pressure. *Nature*, 478:229–232, 2011.

[98] S Burggraf, H Fricke, A Neuner, J Kristjansson, P Rouvier, L Mandelco, C Woese, and K Stetter. *Methanococcus igneus* sp. nov., a novel hyperthermophilic methanogen from a shallow submarine hydrothermal system. *Systematic and Applied Microbiology*, 13:263–269, 1990.

[99] Ralph J Cicerone and Ronald S Oremland. Biogeochemical aspects of atmospheric methane. *Global Biogeochemical Cycles*, 2(4):299–327, 1988.

[100] C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H.-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. Complete genome sequence of the methanogenic archaeon, methanococcus jannaschii. *Science*, 273(5278):1058–1073, Aug 1996.

[101] J.G. Ferry and C.H. House. The stepwise evolution of early life driven by energy conservation. *Molecular Biology and Evolution*, 23(6):1286–1292, 2006.

[102] P.D. Browne and H. Cadillo-Quiroz. Contribution of Transcriptomics to Systems-Level Understanding of Methanogenic *Archaea*. *Archaea*, 2013.

[103] A. Jain, R. Liu, B. Ramani, E. Arauz, Y. Ishitsuka, K. Ragunathan, J. Park, J. Chen, Y.K. Xiang, and T. Ha. Probing cellular protein complexes using single-molecule pull-down. *Nature*, 473(7348):484–488, May 2011.

[104] N.R. Buan and W.W. Metcalf. Methanogenesis by *Methanosarcina acetivorans* involves two structurally and functionally distinct classes of heterodisulfide reductase. *Molecular Microbiology*, 75(4):843–853, 2010.

[105] K. R. Sowers, J. E. Boone, and R. P. Gunsalus. Disaggregation of Methanosarcina spp. and Growth as Single Cells at Elevated Osmolarity. *Applied and Environmental Microbiology*, 59(11):3832–3839, Nov 1993.

[106] William W Metcalf, Jun-Kai Zhang, Xun Shi, and Ralph S Wolfe. Molecular, genetic, and biochemical characterization of the serC gene of *Methanosarcina barkeri* Fusaro. *Journal of Bacteriology*, 178(19):5797–5802, 1996.

[107] W Metcalf, J Zhang, E Apolinario, K Sowers, and R Wolfe. A genetic system for archaea of the genus methanosarcina: liposome-mediated transformation and construction of shuttle vectors. *Proceedings of the National Academy of Sciences of the United States of America*, 94(6):2626–2631, 1997.

[108] Matthew Pritchett, Jun Zhang, and William Metcalf. Development of a markerless genetic exchange method for methanosarcina acetivorans C2A and its use in construction of new genetic tools for methanogenic archaea. *Applied and Environmental Microbiology*, 70(3):1425–1433, 2004.

[109] William W Metcalf, Jun Kai Zhang, and Ralph S Wolfe. An anaerobic, intrachamber incubator for growth of methanosarcina spp. on methanol-containing solid media. *Applied and Environmental Microbiology*, 64(2):768–770, 1998.

[110] K Datsenko and B Wanner. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6640–6645, 2000.

[111] Frank Stewart, Elizabeth Ottesen, and DeLong, Edward. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *The ISME Journal*, 4(7):896–907, 2010.

[112] R. McClure, D. Balasubramanian, Y. Sun, M. Bobrovskyy, P. Sumby, C. A. Genco, C. K. Vanderpool, and B. Tjaden. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research*, 41(14):e140, Aug 2013.

[113] Ankur Jain, Ruijie Liu, Yang Xiang, and Taekjip Ha. Single-molecule pull-down for studying protein interactions. *Nature Protocols*, 7(3):445–452, 2012.

[114] Wolfgang Grabarse, Felix Mahlert, Seigo Shima, Rudolf K Thauer, and Ulrich Ermler. Comparison of three methyl-coenzyme M reductases from phylogenetically distant organisms: unusual amino acid modification, conservation and adaptation. *Journal of Molecular Biology*, 303, 2000.

[115] S.T. Yang and M.R. Okos. Kinetic Study and Mathematical Modeling of Methanogenesis of Acetate Using Pure Cultures of Methanogens. *Biotechnology and Bioengineering*, 30:661–667, 1987.

[116] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI: a COmplex PAthway SImulator. *Bioinformatics*, 22:3067–3074, 2006.

[117] C. Ingram-Smith, A. Gorrell, S.H. Lawrence, P. Iyer, K. Smith, and J.G. Ferry. Characterization of the Acetate Binding Pocket in the *Methanosarcina thermophila* Acetate Kinase. *Journal of Biological Chemistry*, 187(7):2386–2394, 2005.

[118] S.H. Lawrence A. Gorrell and J.G. Ferry. Structural and Kinetic Analyses of Arginine Residues in the Active Site of the Acetate Kinase from *Methanosarcina thermophila*. *Journal of Biological Chemistry*, 280(11):10731–10742, 2005.

[119] S.H. Lawrence, K.B. Luther, H. Schindelin, and J.G. Ferry. Structural and Functional Studies Suggest a Catalytic Mechanism for the Phosphotransacetylase from *Methanosarcina thermophila*. *Journal of Bacteriology*, 188(3):1143–1154, 2006.

[120] M.E. Rasche, K.S. Smith, and J.G. Ferry. Identification of cysteine and arginine residues essential for the phosphotransacetylase from *Methanosarcina thermophila*. *Journal of Bacteriology*, 179(24):7712–7717, 1997.

[121] D.A. Grahame and T.C. Stadtman. Carbon monoxide dehydrogenase from *Methanosarcina barkeri*. Disaggregation, purification, and physicochemical properties of the enzyme. *Journal of Biological Chemistry*, 262(8):3706–3712, 1987.

[122] R.I.L. Eggen, R. van Kranenburg, A.J.M. Vriesema, A.C.M. Geerling, W.R. Hagen M.F.J.M. Verhagen, and W.M. de Vos. Carbon Monoxide Dehydrogenase from *Methanosarcina frisia* Gö1: CHARACTERIZATION OF THE ENZYME AND THE REGULATED EXPRESSION OF TWO OPERON-LIKE *cdh* GENE CLUSTERS. *Journal of Biological Chemistry*, 271(24):14256–14263, 1996.

[123] Matthew A. Pritchett and William W. Metcalf. Genetic, physiological and biochemical characterization of multiple methanol methyltransferase isozymes in methanosarcina acetivorans c2a. *Molecular Microbiology*, 56(5):1183–1194, mar 2005.

[124] K. Ma and R.K. Thauer. Purification and properties of N5, N10-methylenetetrahydromethanopterin reductase from *Methanobacterium thermoautotrophicum* (strain Marburg). *European Journal of Biochemistry*, 191(1):187–193, 1991.

[125] B.W. te Brömmelstroet, W.J. Geerts, J.T. Keltjens, C. van der Drift, and G.D. Vogels. Purification and properties of 5,10-methylenetetrahydromethanopterin dehydrogenase and 5,10-methylenetetrahydromethanopterin reductase, two coenzyme F420-dependent enzymes, from Methanosarcina barkeri. *Biochimica et Biophysica Acta*, 1079:293–302, 1991.

[126] S. Shima and R.K. Thauer. Tetrahydromethanopterin-specific enzymes from *Methanopyrus kandleri*. *Methods in Enzymology*, 331:317–353, 2001.

[127] B.W. te Brömmelstroet, C.M. Hensgens, W.J. Geerts, J.T. Keltjens, C. van der Drift, and G.D. Vogels. Purification and properties of 5,10-methenyltetrahydromethanopterin cyclohydrolase from *Methanosarcina barkeri*. *Journal of Bacteriology*, 172(2):564–571, 1990.

[128] M. Karrasch, G. Böruer, M. Enssle, and R.K. Thauer. The molybdoenzyme formylmethanofuran dehydrogenase from *Methanosarcina barkeri* contains a pterin cofactor. *European Journal of Biochemistry*, 194(2):367–372, 1990.

[129] P.E. Jablonski and J.G. Ferry. Purification and properties of methyl coenzyme M methylreductase from acetate-grown *Methanosarcina thermophila*. *Journal of Bacteriology*, 173(8):2481–2487, 1991.

[130] E. Murakami, U. Deppenmeier, and S.W. Ragsdale. Characterization of the Intramolecular Electron Transfer Pathway from 2-Hydroxyphenazine to the Heterodisulfide Reductase from *Methanosarcina thermophila*. *Journal of Biological Chemistry*, 276(4):2432–2439, 2001.

[131] R. Iino, R. Hasegawa, K.V. Tabata, and H. Noji. Mechanism of Inhibition by C-terminal α-Helices of the ϵ Subunit of *Escherichia coli* FoF1-ATP Synthase. *Journal of Biological Chemistry*, 284(26):17457–17464, 2009.

[132] I. Schomburg, A. Chang, S. Placzek, C. Söhngen, M. Rother, M. Lang, C. Munaretto, S. Ulas, M. Stelzer, A. Grote, M. Scheer, and D. Schomburg. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, 41:764–772, 2013.

[133] R.K. Thauer. Biochemistry of methanogenesis: a tribute to Marjory Stephenson. *Microbiology*, 144:2377–2406, 1998.

[134] R. B. Opulencia, A. Bose, and W. W. Metcalf. Physiology and Posttranscriptional Regulation of Methanol:Coenzyme M Methyltransferase Isozymes in *Methanosarcina acetivorans* C2A. *Journal of Bacteriology*, 191(22):6928–6935, Sep 2009.

[135] R.K. Thauer, K. Jungermann, and K. Decker. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriology*, 1(41):100–180, 1977.

[136] M. W. Peck. Changes in concentrations of coenzyme F420 analogs during batch growth of *Methanosarcina barkeri* and *Methanosarcina mazei*. *Applied and Environmental Microbiology*, 55(4):940–945, Apr 1989.

[137] Boguslaw Obara, Mark Roberts, Judith Armitage, and Vicente Grau. Bacterial cell identification in differential interference contrast microscopy images. *BMC Bioinformatics*, 14, 2013.

[138] Anne Carpenter, Thouis Jones, Michael Lamprecht, Colin Clarke, In Kang, Ola Friman, David Guertin, Joo Chang, Robert Lindquist, Jason Moffat, Polina Golland, and David Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10), 2006.

[139] C. Wrede, U. Walbaum, A. Ducki, I. Heieren, and M. Hoppert. Localization of Methyl-Coenzyme M Reductase as Metabolic Marker for Diverse Methanogenic Archaea. *Archaea*, 2013.

[140] L. Rohlin and R.P. Gunsalus. Carbon-dependent control of electron transfer and central carbon pathway genes for methane biosynthesis in the Archaean, *Methanosarcina acetivorans* C2A. *BMC Microbiology*, 10:62, 2010.

[141] U Ermler, W Grabarse, S Shima, M Goubeaud, and R Thauer. Crystal structure of methyl-coenzyme m reductase: the key enzyme of biological methane formation. *Science*, 278(5342):1457–1462, 1997.

[142] B.E. Krenn, H.S. van Walraven, M.J.C. Scholts, and R. Kraayenhof. Modulation of the proton-translocation stoichiometry of H^+-ATP syntheses in two phototophic prokaryotes by external pH. *Biochemistry Journal*, 294:705–709, 1993.

[143] E. Oelgeschlager and M. Rother. *In vivo* role of three fused corrinoid/methyl transfer proteins in *methanosarcina acetivorans*. *Molecular Microbiology*, 72(5):1260–1272, 2009.

[144] K. Veit, C. Ehlers, A. Ehrenreich, K. Salmon, R. Hovey, R.P. Gunsalus, U. Deppenmeier, and R.A. Schmitz. Global transcriptional analysis of *Methanosarcina mazei* strain Gö1 under different nitrogen availabilities. *Molecular Genetics and Genomics*, 276(1):41–55, 2006.

[145] K. Veit, C. Ehlers, and R.A. Schmitz. Effects of Nitrogen and Carbon Sources on Transcription and Soluble Methyltransferases in *Methanosarcina mazei* Strain Gö1. *Journal of Bacteriology*, 187(17):6147–6154, 2005.

[146] K. Weidenbach, C. Ehlers, J. Kock, and R.A. Schmitz. NrpRII mediates contacts between NrpRI and general transcription factors in the archaeaon *Methanosarcina mazei*, Gö1. *FEBS Journal*, 277(21):4389–4411, 2010.

[147] T.J. Lie, J.A. Dodsworth, D.C. Nickel, and J.A. Leigh. Diverse homologues of the archaeal repressor NrpR function similarly in nitrogen regulation. *FEMS Microbiology Letters*, 271(2):281–288, 2007.

[148] D. Jager, C. M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, and R. A. Schmitz. Deep sequencing analysis of the *Methanosarcina mazei* Go1 transcriptome in response to nitrogen availability. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21878–21882, dec 2009.

[149] C.E. Isom, J.L. Turner, D.J. Lessner, and E.A. Karr. Redox-sensitive DNA binding by homodimeric *Methanosarcina acetivorans* MsvR is modulated by cysteine residues. *BMC Microbiology*, 13(163), 2013.

[150] D. Jager, S. R. Pernitzsch, A. S. Richter, R. Backofen, C. M. Sharma, and R. A. Schmitz. An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains. *Nucleic Acids Research*, 40(21):10964–10979, Sep 2012.

[151] G. Endo and S. Silver. CadC, the Transcriptional Regulatory Protein of the Cadmium Resistance System of *Staphylococcus aureaus* Plasmid pI258. *Journal of Bacteriology*, 177(15):4437–4441, 1995.

[152] E. Lira-Silva, M.G. Santiago-Martinez, V. Hernández-Juárez, R. Garcia-Contreras, R. Moreno-Sánchez, and R. Jasso-Chávez. Activation of Methanogenesis by Cadmium in the Marine Archaeon *Methanosarcina acetivorans*. *PLoS ONE*, 7(11), 2012.

[153] E. Lira-Silva, M.G. Santiago-Martinez, R. Garcia-Contreras, A. Zepeda-Rodríguez, A. Marín-Hernández, R. Moreno-Sánchez, and R. Jasso-Chávez. Cd^{2+} resistance mechanisms in *Methanosarcina acetivorans* involve the increase in coenzyme M content and induction of biofilm synthesis. *Environmental Microbiology Reports*, 2013.

[154] L. Gerosa, K. Kochanowski, M. Heinemann, and U. Sauer. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Molecular Systems Biology*, 9:658, 2013.

[155] John A. Cole, Michael J. Hallock, Piyush Labhsetwar, Joseph R. Peterson, John E. Stone, and Zaida Luthey-Schulten. Stochastic simulations of cellular processes: From single cells to colonies. In Andres Kriete and Roland Eils, editors, *Computational Systems Biology*, chapter 13, pages 277–293. Academic Press, 2 edition, Nov 2013.

[156] J. Fei, D. Singh, Q. Zhang, S. Park, D. Balasubramanian, I. Golding, C. K. Vanderpool, and T. Ha. Determination of in vivo target search kinetics of regulatory noncoding RNA. *Science*, 347(6228):1371–1374, mar 2015.

[157] A. T. Rogowska, O. Puchta, A.M. Czarnecka, A. Kaniak, P.P. Stepien, and P. Golik. Balance between Transcription and RNA Degradation Is Vital for *Saccharomyces cerevisiae* Mitochondria: Reduced Transcription Rescues the Phenotype of Deficient RNA Degradation. *Molecular Biology of the Cell*, 17(3):1184–1193, Dec 2005.

[158] Y. Chiba, K. Mineta, M. Y. Hirai, Y. Suzuki, S. Kanaya, H. Takahashi, H. Onouchi, J. Yamaguchi, and S. Naito. Changes in mRNA Stability Associated with Cold Stress in Arabidopsis Cells. *Plant and Cell Physiology*, 54(2):180–194, Dec 2012.

[159] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, S. Lin-Chao, and S. N. Cohen. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15):9697–9702, jul 2002.

[160] D. W. Selinger, R. M. Saxena, K. J. Cheung, G. M. Church, and C. Rosenow. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Research*, 13(2):216–223, Feb 2003.

[161] T. Esquerré, S. Laguerre, C. Turlan, A. J. Carpousis, L. Girbal, and M. Cocaign-Bousquet. Dual role of transcription and transcript stability in the regulation of gene expression in *Escherichia coli* cells cultured on glucose at different growth rates. *Nucleic Acids Research*, 42(4):2460–2472, Nov 2013.

[162] C. Dressaire, F. Picard, E. Redon, P. Loubière, I. Queinnec, L. Girbal, and M. Cocaign-Bousquet. Role of mRNA Stability during Bacterial Adaptation. *PLoS ONE*, 8(3):e59059, Mar 2013.

[163] Thomas Esquerré, Annick Moisan, Hélène Chiapello, Liisa Arike, Raivo Vilu, Christine Gaspin, Muriel Cocaign-Bousquet, and Laurence Girbal. Genome-wide investigation of mRNA lifetime determinants in *Escherichia coli* cells cultured at different growth rates. *BMC Genomics*, 16(1), Apr 2015.

[164] T. R. Rustad, K. J. Minch, W. Brabant, J. K. Winkler, D. J. Reiss, N. S. Baliga, and D. R. Sherman. Global analysis of mRNA stability in *Mycobacterium tuberculosis*. *Nucleic Acids Research*, 41(1):509–517, Nov 2012.

[165] G. Hambraeus, C. von Wachenfeldt, and L. Hederstedt. Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. *Molecular Genetics and Genomics*, 269(5):706–714, Jul 2003.

[166] S.M. Kristoffersen, C. Haase, M.R. Weil, K.D. Passalacqua, F. Niazi, S.K. Hutchison, B. Desany, A.B. Kolstø, N.J. Tourasse, T.D. Read, and O. Økstad. Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a gram-positive bacterium. *Genome Biology*, 13(4):R30, 2012.

[167] E. Bini, V. Dikshit, K. Dirksen, M. Drozda, and P. Blum. Stability of mRNA in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *RNA*, 8(9):1129–1136, Sep 2002.

[168] A.F. Andersson, M. Lundgren, S. Eriksson, M. Rosenlund, R. Bernander, and P. Nilsson. Global analysis of mRNA stability in the archaeon *Sulfolobus*. *Genome Biology*, 7(10):R99, 2006.

[169] S. Hundt, A. Zaigler, C. Lange, J. Soppa, and G. Klug. Global Analysis of mRNA Decay in *Halobacterium salinarum* NRC-1 at Single-Gene Resolution Using DNA Microarrays. *Journal of Bacteriology*, 189(19):6936–6944, Jul 2007.

[170] J. Zhang and G. J. Olsen. Messenger RNA processing in *Methanocaldococcus (Methanococcus) jannaschii*. *RNA*, 15(10):1909–1916, Oct 2009.

[171] Y. Wang, C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America*, 99(9):5860–5865, Apr 2002.

[172] J. Grigull, S. Mnaimneh, J. Pootoolal, M. D. Robinson, and T. R. Hughes. Genome-Wide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors. *Molecular and Cellular Biology*, 24(12):5534–5547, May 2004.

[173] S. E. Munchel, R. K. Shultzaberger, N. Takizawa, and K. Weis. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Molecular Biology of the Cell*, 22(15):2787–2795, Jun 2011.

[174] J.V. Geisberg, Z. Moqtaderi, X. Fan, F. Ozsolak, and K. Struhl. Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. *Cell*, 156(4):812–824, Feb 2014.

[175] D. H. Rothman, G. P. Fournier, K. L. French, E. J. Alm, E. A. Boyle, C. Cao, and R. E. Summons. Methanogenic burst in the end-Permian carbon cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 111(15):5462–5467, Mar 2014.

[176] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(24):9743–9748, May 2009.

[177] J. Li, L. Qi, Y. Guo, L. Yue, Y. Li, W. Ge, J. Wu, W. Shi, and X. Dong. Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon *Methanolobus psychrophilus*. *Scientific Reports*, 5:9209, 2015.

[178] Ralph S. Wolfe. Chapter one - techniques for cultivating methanogens. In Amy C. Rosenzweig and Stephen W. Ragsdale, editors, *Methods in Methane Metabolism, Part A*, volume 494 of *Methods in Enzymology*, pages 1 – 22. Academic Press, 2011.

[179] F.K. Stewart, E.A. Ottesen, and E.F. DeLong. Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *The ISME Journal*, 4(7):896–907, Mar 2010.

[180] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1):D36–D42, Nov 2012.

[181] K.J. Millman and M. Aivazis. Python for Scientists and Engineers. *Computing in Science & Engineering*, 13(2):9–12, Mar 2011.

[182] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, and I.L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[183] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292, May 2004.

[184] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–i28, Jul 2007.

[185] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

[186] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, Oct 2011.

[187] M.I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), Dec 2014.

[188] M. N. Price, K. H. Huang, E. J. Alm, and A. P. Arkin. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research*, 33(3):880–892, Feb 2005.

[189] B. Taboada, R. Ciria, C. E. Martinez-Guerrero, and E. Merino. ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Research*, 40(D1):D627–D631, Nov 2011.

[190] X. Mao, Q. Ma, C. Zhou, X. Chen, H. Zhang, J. Yang, F. Mao, W. Lai, and Y. Xu. DOOR 2.0: presenting operons and their functions through dynamic and integrated views. *Nucleic Acids Research*, 42(D1):D654–D659, Nov 2013.

[191] Matthew N Benedict, James R Henriksen, William W Metcalf, Rachel J Whitaker, and Nathan D Price. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics*, 15(1):8, 2014.

[192] Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638, feb 2016.

[193] A.S. Blazier and J.A. Papin. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*, 3, 2012.

[194] A. M. Guss, B. Mukhopadhyay, J. K. Zhang, and W. W. Metcalf. Genetic analysis of mch mutants in two Methanosarcina species demonstrates multiple roles for the methanopterin-dependent C-1 oxidation/reduction pathway and differences in H(2) metabolism between closely related species. *Molecular Microbiology*, 55(6):1671–1680, Mar 2005.

[195] K. R. Sowers, S. F. Baron, and J. G. Ferry. *Methanosarcina acetivorans* sp. nov., an Acetotrophic Methane-Producing Bacterium Isolated from Marine Sediments. *Applied and Environmental Microbiology*, 47(5):971–978, May 1984.

[196] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1):D261–D269, Nov 2014.

[197] K. Makarova, Y. Wolf, and E. Koonin. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life*, 5(1):818–840, Mar 2015.

[198] H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1):D377–D386, Nov 2012.

[199] K.L. Anderson, E.E. Apolinario, S.R. MacAuley, and K.R. Sowers. A 5′ Leader Sequence Regulates Expression of Methanosarcinal CO Dehydrogenase/Acetyl Coenzyme A Synthase. *Journal of Bacteriology*, 191(22):7123–7128, 2009.

[200] A. Sauerwald, W. Zhu, T. A. Major, H. Roy, S. Palioura, D. Jahn, W. B. Whitman, J. R. Yates 3rd, M. Ibba, and D. Söll. RNA-Dependent Cysteine Biosynthesis in Archaea. *Science*, 307(5717):1969–1972, Mar 2005.

[201] D. Miller, Y. Wang, H. Xu, K. Harich, and R.H. White. Biosynthesis of the 5-(Aminomethyl)-3-furanmethanol Moiety of Methanofuran. *Biochemistry*, 53(28):4635–4647, Jul 2014.

[202] Y. Wang, H. Xu, K.C. Harich, and R.H. White. Identification and Characterization of a Tyramine–Glutamate Ligase (MfnD) Involved in Methanofuran Biosynthesis. *Biochemistry*, 53(39):6220–6230, Oct 2014.

[203] Y. Wang, M.K. Jones, H. Xu, W.K. Ray, and R.H. White. Mechanism of the Enzymatic Synthesis of 4-(Hydroxymethyl)-2-furancarboxaldehyde-phosphate (4-HFC-P) from Glyceraldehyde-3-phosphate Catalyzed by 4-HFC-P Synthase. *Biochemistry*, 54(19):2997–3008, May 2015.

[204] Y. Wang, H. Xu, and R. H. White. Identification of the Final Two Genes Functioning in Methanofuran Biosynthesis in *Methanocaldococcus jannaschii*. *Journal of Bacteriology*, 197(17):2850–2858, Sep 2015.

[205] K. Isobe, T. Ogawa, K. Hirose, T. Yokoi, T. Yoshimura, and H. Hemmi. Geranylgeranyl reductase and ferredoxin from *methanosarcina acetivorans* are required for the synthesis of fully reduced archaeal membrane lipid in escherichia coli cells. *Journal of Bacteriology*, 196(2):417–423, nov 2013.

[206] Takeshi Mori, Keisuke Isobe, Takuya Ogawa, Tohru Yoshimura, and Hisashi Hemmi. A phytoene desaturase homolog gene from the methanogenic archaeon *methanosarcina acetivorans* is responsible for hydroxyarchaeol biosynthesis. *Biochemical and Biophysical Research Communications*, 466(2):186–191, oct 2015.

[207] Takuya Ogawa, Koh ichi Emi, Kazushi Koga, Tohru Yoshimura, and Hisashi Hemmi. A cis-prenyltransferase from *methanosarcina acetivorans* catalyzes both head-to-tail and nonhead-to-tail prenyl condensation. *FEBS Journal*, 283(12):2369–2383, may 2016.

[208] K. R. Sowers and R. P. Gunsalus. Halotolerance in Methanosarcina spp: Role of N-Acetyl-beta-Lysine, alpha-Glutamate, Glycine Betaine,

and K+ as compatible solutes for Osmotic Adaptation. *Applied and Environmental Microbiology*, 61(12):4382–4388, Dec 1995.

[209] Michel Geovanni Santiago-Martínez, Rusely Encalada, Elizabeth Lira-Silva, Erika Pineda, Juan Carlos Gallardo-Pérez, Marco Antonio Reyes-García, Emma Saavedra, Rafael Moreno-Sánchez, Alvaro Marín-Hernández, and Ricardo Jasso-Chávez. The nutritional status of *methanosarcina acetivorans* regulates glycogen metabolism and gluconeogenesis and glycolysis fluxes. *FEBS Journal*, 283(10):1979–1999, apr 2016.

[210] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, Dec 2010.

[211] Z.A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N.E. Lewis, and B.Ø. Palsson. Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLoS Computational Biology*, 11(8):e1004321, Aug 2015.

[212] Y. Cao, J. Li, N. Jiang, and X. Dong. Mechanism for Stabilizing mRNAs Involved in Methanol-Dependent Methanogenesis of Cold-Adaptive *Methanosarcina mazei* zm-15. *Applied and Environmental Microbiology*, 80(4):1291–1298, Dec 2013.

[213] C. Kratzer, P. Carini, R. Hovey, and U. Deppenmeier. Transcriptional Profiling of Methyltransferase Genes during Growth of *Methanosarcina mazei* on Trimethylamine. *Journal of Bacteriology*, 191(16):5108–5115, Jun 2009.

[214] Nicholas D Youngblut, Joseph S Wirth, James R Henriksen, Maria Smith, Holly Simon, William W Metcalf, and Rachel J Whitaker. Genomic and phenotypic differentiation among methanosarcina mazei populations from columbia river sediment. *The ISME Journal*, 9(10):2191–2205, Mar 2015.

[215] D. R. Boone, I. M. Mathrani, Y. Liu, J. A. G. F. Menaia, R. A. Mah, and J. E. Boone. Isolation and Characterization of *Methanohalophilus portucalensis* sp. nov. and DNA Reassociation Study of the Genus Methanohalophilus. *International Journal of Systematic Bacteriology*, 43(3):430–437, Jul 1993.

[216] D. Wilson, V. Charoensawan, S.K. Kummerfeld, and S.A. Teichmann. DBD taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, 36(suppl_1):D88–92, 2008.

[217] J.L. Catlett, A.M. Ortiz, and N.R. Buan. Rerouting Cellular Electron Flux To Increase the Rate of Biological Methane Production. *Applied and Environmental Microbiology*, 81(19):6528–6537, Jul 2015.

[218] S. H. Yoon, S. Turkarslan, D. J. Reiss, M. Pan, J. A. Burn, K. C. Costa, T. J. Lie, J. Slagel, R. L. Moritz, M. Hackett, J. A. Leigh, and N. S. Baliga. A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen. *Genome Research*, 23(11):1839–1851, Oct 2013.

[219] F.C. Neidhardt, R. Curtiss III, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, M. Schaechter, and H.E. Umbarger, editors. *Escherichia coli and Salmonella: cellular and molecular biology*. ASM Press, Washington, D.C., 2 edition, 1996.

[220] Alexander Zaigler, Stephan C. Schuster, and Jörg Soppa. Construction and usage of a onefold-coverage shotgun DNA microarray to characterize the metabolism of the archaeon haloferax volcanii. *Molecular Microbiology*, 48(4):1089–1105, may 2003.

[221] Nahum Sonenberg and Alan G. Hinnebusch. Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. *Cell*, 136(4):731–745, feb 2009.

[222] Christian Lange, Alexander Zaigler, Mathias Hammelmann, Jens Twellmeyer, Günter Raddatz, Stephan C. Schuster, Dieter Oesterhelt, and Jörg Soppa. Genome-wide analysis of growth phase-dependent translational and transcriptional regulation in halophilic archaea. *BMC Genomics*, 8(1):1–16, 2007.

[223] Mariam Brenneis and Jörg Soppa. Regulation of translation in haloarchaea: 5′- and 3′-utrs are essential and have to functionally interact *In Vivo*. *PLoS ONE*, 4(2):1–8, 02 2009.

[224] R. Jasso-Chávez, M.G. Santiago-Martínez, E. Lira-Silva, E. Pineda, A. Zepeda-Rodríguez, J. Belmont-Díaz, R. Encalada, E. Saavedra, and R. Moreno-Sánchez. Air-Adapted *Methanosarcina acetivorans* Shows High Methane Production and Develops Resistance against Oxygen Stress. *PLoS ONE*, 10(2):e0117331, Feb 2015.

[225] R. Schuetz, N. Zamboni, M. Zampieri, M. Heinemann, and U. Sauer. Multidimensional Optimality of Microbial Metabolism. *Science*, 336(6081):601–604, May 2012.

[226] F.Y. Edgeworth. XXII. on a new method of reducing observations relating to several quantities. *Philosophical Magazine Series 5*, 25(154):184–191, Mar 1888.

[227] F. Seyednasrollah, A. Laiho, and L. L. Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70, Dec 2013.

[228] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C.E. Mason, N.D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013.

[229] S. Sundararaj. The CyberCell database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of *escherichia coli*. *Nucleic Acids Research*, 32(90001):293D–295, Jan 2004.

[230] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. COBRApy: COnstraints-based reconstruction and analysis for python. *BMC Systems Biology*, 7(1):74, 2013.

[231] Y. Liu, L. L. Beer, and W. B. Whitman. Methanogens: a window into ancient sulfur metabolism. *Trends in Microbiology*, 20(5):251–258, May 2012.

[232] B. Hao, W. Gong, T. K. Ferguson, C. M. James, J. A. Krzycki, and M. K. Chan. A New UAG-Encoded Residue in the Structure of a Methanogen Methyltransferase. *Science*, 296(5572):1462–1466, May 2002.

[233] A. Mahapatra, A. Patel, J. A. Soares, R. C. Larue, J. K. Zhang, W. W. Metcalf, and J. A. Krzycki. Characterization of a *Methanosarcina acetivorans* mutant unable to translate UAG as pyrrolysine. *Molecular Microbiology*, 59(1):56–66, Jan 2006.

[234] M. A. Gaston, L. Zhang, K. B. Green-Church, and J. A. Krzycki. The complete biosynthesis of the genetically encoded amino acid pyrrolysine. *Nature*, 471:647–650, Mar 2011.

[235] M. A. Pritchett and W. W. Metcalf. Genetic, physiologic and biochemical characterization of multiple methanol methyltransferase isozymes in *methanosarcina acetiverans* c2a. *Molecular Microbiology*, 56(5):1183–1194, 2005.

[236] K. R. Sowers, M. J. Nelson, and J. G. Ferry. Growth of acetotrophic, methane-producing bacteria in a ph auxostat. *Current Microbiology*, 11:227–229, 1984.

[237] H. Summer. Improved aproach for transferring and cultivating *methanosarcina acetivorans*. *Letters in Applied Microbiology*, 48(6):786–789, 2009.

[238] E. Oelgeschlager and M. Rother. Influence of carbon monoxide on metabolite formation in *methanosarcina acetivorans*. *FEMS Microbiology Letters*, 292(2):254–260, 2009.

[239] D. J. Lessner, L. Lhu, C. S. Wahal, and J. G. Ferry. An engineered methanogenic pathway derived from the domains *bacteria* and *archaea*. *MBio*, 1(5):1–4, 2010.

[240] E. Heinie-Dobbernack, S.M. Schoberth, and H. Sahm. Relationship of Intracellular Coenzyme F420 Content to Growth and Metabolic Activity of *Methanobacterium bryantii* and *Methanosarcina barkeri*. *Applied and Environmental Microbiology*, 54(2):454–459, Feb 1988.

[241] L. Baresi and R. S. Wolfe. Levels of coenzyme F420, coenzyme M, hydrogenase, and methylcoenzyme M methylreductase in acetate-grown Methanosarcina. *Applied and Environmental Microbiology*, 41(2):388–391, Feb 1981.

[242] L.G.M. Gorris and C. van der Drift. Methanogenic cofactors in pure cultures of methanogens in relation to substrate utilization. In H. C. Dubourguier, G. Albagnac, J. Montreuil, C. Romond, P. Sautiere, and J. Guillaume, editors, *Biology of anaerobic bacteria*, pages 144–150. Elsevier Science Publishing, Inc., Amsterdam, 1988.

[243] W. E. Balch, L. J. Magrum, G. E. Fox, R. S. Wolfe, and C. R. Woese. An ancient divergence among the bacteria. *Journal of Molecular Biology*, 9(4):305–311, Aug 1977.

[244] C. R. Woese, L. J. Magrum, and G. E. Fox. Archaebacteria. *Journal of Molecular Biology*, 11(3):245–251, Aug 1978.

[245] L. J. Magrum, K. R. Luehrsen, and C. R. Woese. Are extreme halophiles actually "bacteria"? *Journal of Molecular Biology*, 11(1):1–8, May 1978.

[246] C. R. Woese and R. Gupta. Are archaebacteria merely derived 'prokaryotes'? *Nature*, 289(5793):95–96, Jan 1981.

[247] Charles E. Robertson, J. Kirk Harris, John R. Spear, and Norman R. Pace. Phylogenetic diversity and ecology of environmental archaea. *Current Opinion in Microbiology*, 8(6):638–642, Dec 2005.

[248] Gary J. Olsen and Carl R. Woese. Archaeal genomics: An overview. *Cell*, 89(7):991–994, Jun 1997.

[249] Sonja-Verena Albers, Patrick Forterre, David Prangishvili, and Christa Schleper. The legacy of carl woese and wolfram zillig: from phylogeny to landmark discoveries. *Nature Reviews Microbiology*, 11(10):713–719, Sep 2013.

[250] Brandon K Swan and David L Valentine. *Diversity of Archaea*. John Wiley and Sons, Ltd, 2001.

[251] Michael T. Madigan, John M. Martinko, Kelly S. Bender, Daniel H. Buckley, David A. Stahl, and Thomas Brock. *Brock Biology of Microorganisms*. Pearson, 2014.

[252] David L. Valentine. Adaptations to energy stress dictate the ecology and evolution of the archaea. *Nature Reviews Microbiology*, 5(4):316–323, Mar 2007.

[253] J. G. Elkins, M. Podar, D. E. Graham, K. S. Makarova, Y. Wolf, L. Randau, B. P. Hedlund, C. Brochier-Armanet, V. Kunin, I. Anderson, A. Lapidus, E. Goltsman, K. Barry, E. V. Koonin, P. Hugenholtz, N. Kyrpides, G. Wanner, P. Richardson, M. Keller, and K. O. Stetter. A korarchaeal genome reveals insights into the evolution of the archaea. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23):8102–8107, Jun 2008.

[254] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269(5223):496–512, Jul 1995.

[255] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, Mar 2003.

[256] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard Ø Palsson, and Markus J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nature Protocols*, 2(3):727–738, Mar 2007.

[257] Jan Schellenberger, Richard Que, Ronan M T Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, Joseph Kang, Daniel R Hyduke, and Bernhard Ø Palsson. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox v2.0. *Nature Protocols*, 6(9):1290–1307, Aug 2011.

[258] Di Wu, Qin Wang, Rajeev S. Assary, Linda J. Broadbelt, and Goran Krilov. A computational approach to design and evaluate enzymatic reaction pathways: Application to 1-butanol production from pyruvate. *Journal of Chemical Information and Modeling*, 51(7):1634–1647, Jul 2011.

[259] D. A. Pertusi, A. E. Stine, L. J. Broadbelt, and K. E. J. Tyo. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics*, 31(7):1016–1024, Nov 2014.

[260] Miguel A. Campodonico, Barbara A. Andrews, Juan A. Asenjo, Bernhard O. Palsson, and Adam M. Feist. Generation of an atlas for commodity chemical production in escherichia coli and a novel pathway prediction algorithm, GEM-path. *Metabolic Engineering*, 25:140–158, Sep 2014.

[261] Anupam Chowdhury and Costas D. Maranas. Designing overall stoichiometric conversions and intervening metabolic reactions. *Scientific Reports*, 5:16009, Nov 2015.

[262] Isabel Rocha, Paulo Maia, Pedro Evangelista, Paulo Vilaça, Simão Soares, José P Pinto, Jens Nielsen, Kiran R Patil, Eugénio C Ferreira, and Miguel Rocha. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Systems Biology*, 4(1):45, 2010.

[263] Sridhar Ranganathan, Patrick F. Suthers, and Costas D. Maranas. OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Computational Biology*, 6(4):e1000744, Apr 2010.

[264] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, 84(6):647–657, Dec 2003.

[265] Kai Zhuang, Mounir Izallalen, Paula Mouser, Hanno Richter, Carla Risso, Radhakrishnan Mahadevan, and Derek R Lovley. Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *The ISME Journal*, 5(2):305–316, Jul 2010.

[266] Ali R. Zomorrodi and Costas D. Maranas. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Computational Biology*, 8(2):e1002363, feb 2012.

[267] Ruchir A. Khandelwal, Brett G. Olivier, Wilfred F. M. Röling, Bas Teusink, and Frank J. Bruggeman. Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE*, 8(5):e64567, May 2013.

[268] Saeed Shoaie, Fredrik Karlsson, Adil Mardinoglu, Intawat Nookaew, Sergio Bordel, and Jens Nielsen. Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific Reports*, 3, Aug 2013.

[269] Wilfred F. M. Röling and Peter M. van Bodegom. Toward quantitative understanding on microbial community structure and functioning: a modeling-centered approach using degradation of marine oil spills as example. *Frontiers in Microbiology*, 5, Mar 2014.

[270] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, Oct 2015.

[271] UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, Oct 2014.

[272] Ron Caspi, Tomer Altman, Richard Billington, Kate Dreher, Hartmut Foerster, Carol A. Fulcher, Timothy A. Holland, Ingrid M. Keseler, Anamika Kothari, Aya Kubo, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Quang Ong, Suzanne Paley, Pallavi Subhraveti, Daniel S. Weaver, Deepika Weerasinghe, Peifen Zhang, and Peter D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, 42(D1):D459–D471, Nov 2013.

[273] Rasmus Agren, Liming Liu, Saeed Shoaie, Wanwipa Vongsangnak, Intawat Nookaew, and Jens Nielsen. The RAVEN toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS Computational Biology*, 9(3):e1002980, Mar 2013.

[274] Scott Devoid, Ross Overbeek, Matthew DeJongh, Veronika Vonstein, Aaron A. Best, and Christopher Henry. Automated genome annotation and metabolic model reconstruction in the SEED and model SEED. In *Methods in Molecular Biology*, pages 17–45. Springer Science + Business Media, 2013.

[275] Vinay Satish Kumar, Madhukar S Dasika, and Costas D Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):212, 2007.

[276] Nathan D. Price, Jennifer L. Reed, and Bernhard Ø. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, Nov 2004.

[277] O. Dias, M. Rocha, E. C. Ferreira, and I. Rocha. Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Research*, 43(8):3899–3910, Apr 2015.

[278] Esa Pitkänen, Paula Jouhten, Jian Hou, Muhammad Fahad Syed, Peter Blomberg, Jana Kludas, Merja Oja, Liisa Holm, Merja Penttilä, Juho Rousu, and Mikko Arvas. Comparative genome-scale reconstruction

of gapless metabolic networks for present and ancestral species. *PLoS Computational Biology*, 10(2):e1003465, Feb 2014.

[279] Peter D. Karp, Mario Latendresse, Suzanne M. Paley, Markus Krummenacker, Quang D. Ong, Richard Billington, Anamika Kothari, Daniel Weaver, Thomas Lee, Pallavi Subhraveti, Aaron Spaulding, Carol Fulcher, Ingrid M. Keseler, and Ron Caspi. Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, page bbv079, Oct 2015.

[280] Xueyang Feng, You Xu, Yixin Chen, and Yinjie J Tang. MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *BMC Systems Biology*, 6(1):94, 2012.

[281] Stephan Pabinger, Robert Rader, Rasmus Agren, Jens Nielsen, and Zlatko Trajanoski. MEMOSys: Bioinformatics platform for genome-scale metabolic models. *BMC Systems Biology*, 5(1):20, 2011.

[282] Joost Boele, Brett G Olivier, and Bas Teusink. FAME, the flux analysis and modeling environment. *BMC Systems Biology*, 6(1):8, 2012.

[283] S. Tsoka, D. Simon, and C. A. Ouzounis. Automated metabolic reconstruction for *Methanococcus jannaschii*. *Archaea*, 1(4):223–229, Oct 2004.

[284] M. N. Benedict, M. C. Gonnerman, W. W. Metcalf, and N. D. Price. Genome-scale metabolic reconstruction and hypothesis testing in the methanogenic archaeon methanosarcina acetivorans c2a. *Journal of Bacteriology*, 194(4):855–865, dec 2011.

[285] Joshua J. Hamilton, Montserrat Calixto Contreras, and Jennifer L. Reed. Thermodynamics and h2 transfer in a methanogenic, syntrophic community. *PLoS Computational Biology*, 11(7):e1004364, Jul 2015.

[286] Hadi Nazem-Bokaee, Saratram Gopalakrishnan, James G. Ferry, Thomas K. Wood, and Costas D. Maranas. Assessing methanotrophy and carbon fixation for biofuel production by *Methanosarcina acetivorans*. *Microbial Cell Factories*, 15(1), Jan 2016.

[287] N. Goyal, H. Widiastuti, I. A. Karimi, and Z. Zhou. A genome-scale metabolic model of *Methanococcus maripaludis* S2 for CO 2 capture and conversion to methane. *Molecular BioSystems*, 10(5):1043–1054, 2014.

[288] Nishu Goyal, Mrutyunjay Padhiary, Iftekhar A. Karimi, and Zhi Zhou. Flux measurements and maintenance energy for carbon dioxide utilization by *Methanococcus maripaludis*. *Microbial Cell Factories*, 14(1), Sep 2015.

[289] C. Petitjean, P. Deschamps, P. Lopez-Garcia, and D. Moreira. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom proteoarchaeota. *Genome Biology and Evolution*, 7(1):191–204, Dec 2014.

[290] Orland Gonzalez, Susanne Gronau, Michaela Falb, Friedhelm Pfeiffer, Eduardo Mendoza, Ralf Zimmer, and Dieter Oesterhelt. Reconstruction, modeling & analysis of *Halobacterium salinarum* r-1 metabolism. *Molecular BioSystems*, 4(2):148–159, 2008.

[291] Orland Gonzalez, Susanne Gronau, Friedhelm Pfeiffer, Eduardo Mendoza, Ralf Zimmer, and Dieter Oesterhelt. Systems analysis of bioenergetics and growth of the extreme halophile *Halobacterium salinarum*. *PLoS Computational Biology*, 5(4):e1000332, Apr 2009.

[292] Matthew C. Gonnerman, Matthew N. Benedict, Adam M. Feist, William W. Metcalf, and Nathan D. Price. Genomically and biochemically accurate metabolic reconstruction of *Methanosarcina barkeri* fusaro, iMG746. *Biotechnology Journal*, 8(9):1070–1079, Mar 2013.

[293] Marcin Bizukojc, David Dietz, Jibin Sun, and An-Ping Zeng. Metabolic modelling of syntrophic-like growth of a 1, 3-propanediol producer, *Clostridium butyricum*, and a methanogenic archeon, *Methanosarcina mazei*, under anaerobic conditions. *Bioprocess Biosyst Eng*, 33(4):507–523, Aug 2009.

[294] Orland Gonzalez, Tanja Oberwinkler, Locedie Mansueto, Friedhelm Pfeiffer, Eduardo Mendoza, Ralf Zimmer, and Dieter Oesterhelt. Characterization of growth and metabolism of the haloalkaliphile *Natronomonas pharaonis*. *PLoS Computational Biology*, 6(6):e1000799, Jun 2010.

[295] Thomas Ulas, S. Alexander Riemer, Melanie Zaparty, Bettina Siebers, and Dietmar Schomburg. Genome-scale reconstruction and analysis of the metabolic network in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *PLoS ONE*, 7(8):e43401, Aug 2012.

[296] Q. Li, L. Li, T. Rejtar, D. J. Lessner, B. L. Karger, and J. G. Ferry. Electron transport in the pathway of acetate conversion to methane in the marine archaeon *Methanosarcina acetivorans*. *Journal of Bacteriology*, 188(2):702–710, Dec 2005.

[297] Otto Kandler and Hans Hippe. Lack of peptidoglycan in the cell walls of *Methanosarcina barkeri*. *Archives of Microbiology*, 113(1-2):57–60, 1977.

[298] Patrick F. Suthers, Madhukar S. Dasika, Vinay Satish Kumar, Gennady Denisov, John I. Glass, and Costas D. Maranas. A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. *PLoS Computational Biology*, 5(2):e1000285, Feb 2009.

[299] Miriam Kolog Gulko, Mike Dyall-Smith, Orland Gonzalez, and Dieter Oesterhelt. How do haloarchaea synthesize aromatic amino acids? *PLoS ONE*, 9(9):e107475, Sep 2014.

[300] P. Cabello. Nitrate reduction and the nitrogen cycle in archaea. *Microbiology*, 150(11):3527–3546, Nov 2004.

[301] Jing Zhao, Hong Yu, Jian-Hua Luo, Zhi-Wei Cao, and Yi-Xue Li. Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics*, 7(1):386, 2006.

[302] Z. N. Oltvai, A.-L. Barabási, H. Jeong, B. Tombor, and R. Albert. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[303] Matthew A Oberhardt, Bernhard Ø Palsson, and Jason A Papin. Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, 5, nov 2009.

[304] Caroline B. Milne, Pan-Jun Kim, James A. Eddy, and Nathan D. Price. Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnology Journal*, 4(12):1653–1670, dec 2009.

[305] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, 7(1):501–501, apr 2014.

[306] Tae Yong Kim, Seung Bum Sohn, Yu Bin Kim, Won Jun Kim, and Sang Yup Lee. Recent advances in reconstruction and applications

of genome-scale metabolic models. *Current Opinion in Biotechnology*, 23(4):617–623, aug 2012.

[307] J. L. Reed and B. Ø. Palsson. Thirteen years of building constraint-based in silico models of escherichia coli. *Journal of Bacteriology*, 185(9):2692–2699, may 2003.

[308] Adam M. Feist, Markus J. Herrgård, Ines Thiele, Jennie L. Reed, and Bernhard Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, dec 2008.

[309] J. M. Peregrin-Alvarez. The phylogenetic extent of metabolic enzymes and pathways. *Genome Research*, 13(3):422–427, mar 2003.

[310] Joshua J. Hamilton and Jennifer L. Reed. Identification of functional differences in metabolic networks using comparative genomics and constraint-based models. *PLoS ONE*, 7(4):e34670, apr 2012.

[311] J. M. Monk, P. Charusanti, R. K. Aziz, J. A. Lerman, N. Premyodhin, J. D. Orth, A. M. Feist, and B. O. Palsson. Genome-scale metabolic reconstructions of multiple escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences of the United States of America*, 110(50):20338–20343, nov 2013.

[312] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie I Stefan, Jacky L Snoep, Michael Hucka, Nicolas Le Novère, and Camille Laibe. BioModels database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 4(1):92, 2010.

[313] Christopher S Henry, Matthew DeJongh, Aaron A Best, Paul M Frybarger, Ben Linsay, and Rick L Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9):977–982, aug 2010.

[314] P. D. Karp, S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, P. Kaipa, F. Gilham, A. Spaulding, L. Popescu, T. Altman, I. Paulsen, I. M. Keseler, and R. Caspi. Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 11(1):40–79, dec 2009.

[315] Alexandra M. Schnoes, Shoshana D. Brown, Igor Dodevski, and Patricia C. Babbitt. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5(12):e1000605, dec 2009.

[316] Amrita Pati, Natalia N Ivanova, Natalia Mikhailova, Galina Ovchinnikova, Sean D Hooper, Athanasios Lykidis, and Nikos C Kyrpides. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Methods*, 7(6):455–457, may 2010.

[317] M. S. Poptsova and J. P. Gogarten. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*, 156(7):1909–1917, apr 2010.

[318] Yan Ji, Yixiang Shi, Guohui Ding, and Yixue Li. A new strategy for better genome assembly from very short reads. *BMC Bioinformatics*, 12(1):493, 2011.

[319] José P. Faria, Janaka N. Edirisinghe, James J. Davis, Terrence Disz, Anna Hausmann, Christopher S. Henry, Robert Olson, Ross A. Overbeek, Gordon D. Pusch, Maulik Shukla, Veronika Vonstein, and Alice R. Wattam. Enabling comparative modeling of closely related genomes: example genus brucella. *3 Biotech*, 5(1):101–105, mar 2014.

[320] A. M. Guss, G. Kulkarni, and W. W. Metcalf. Differences in hydrogenase gene expression between methanosarcina acetivorans and methanosarcina barkeri. *Journal of Bacteriology*, 191(8):2826–2833, feb 2009.

[321] U. Deppenmeier. The unique biochemistry of methanogenesis. *Proceedings of the National Academy of Sciences*, 71:223–283, 2002.

[322] D. A. Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 40(D1):D48–D53, dec 2011.

[323] Ramy K Aziz, Daniela Bartels, Aaron A Best, Matthew DeJongh, Terrence Disz, Robert A Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross A Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas

Wilke, and Olga Zagnitko. The RAST server: Rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75, 2008.

[324] L. Li. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189, sep 2003.

[325] D. L. Maeder, I. Anderson, T. S. Brettin, D. C. Bruce, P. Gilna, C. S. Han, A. Lapidus, W. W. Metcalf, E. Saunders, R. Tapia, and K. R. Sowers. The methanosarcina barkeri genome: Comparative analysis with methanosarcina acetivorans and methanosarcina mazei reveals extensive rearrangement within methanosarcinal genomes. *Journal of Bacteriology*, 188(22):7922–7931, Sep 2006.

[326] Robert A. Mah. Isolation and characterization ofMethanococcus mazei. *Current Microbiology*, 3(6):321–326, nov 1980.

[327] Dominique von Klein, Hocine Arab, Horst Völker, and Michael Thomm. Methanosarcina baltica, sp. nov., a novel methanogen isolated from the gotland deep of the baltic sea. *Extremophiles*, 6(2):103–110, jan 2002.

[328] Kai Finster, Yuichi Tanimoto, and Friedhelm Bak. Fermentation of methanethiol and dimethylsulfide by a newly isolated methanogenic bacterium. *Archives of Microbiology*, 157(5):425–430, may 1992.

[329] S.H. Zinder and R.A. Mah. Isolation and characterization of a thermophilic strain of *methanosarcina* unable to use h2–co2 for methanogenesis. *Applied and Environmental Microbiology*, 38(5):996–1008, Nov 1979.

[330] J.P. Touzel, D. Petroff, and G. Albagnac. Isolation and characterization of a new thermophilic methanosarcina, the strain CHTI 55. *Systematic and Applied Microbiology*, 6(1):66–71, jun 1985.

[331] G. Albagnac and J.P. Touzel. *Acetoclastic methanogens in anaerobic digesters*, pages 35–39. Noyes Data Corporation, Park Ridge New Jersey, jun 1987.

[332] T. N. Zhilina and G. A. Zavarzin. Methanosarcina vacuolata sp. nov., a vacuolated methanosarcina. *International Journal of Systematic Bacteriology*, 37(3):281–283, jul 1987.

[333] H. Hippe, D. Caspari, K. Fiebig, and G. Gottschalk. Utilization of trimethylamine and other n-methyl compounds for growth and methane formation by methanosarcina-barkeri. *International Journal of Systematic Bacteriology*, 76(1):494–498, 1979.

[334] P. Scherer, H. Lippert, and G. Wolff. Composition of the major elements and trace elements of 10 methanogenic bacteria determined by inductively coupled plasma emission spectrometry. *Biological Trace Element Research*, 5(3):149–163, jun 1983.

[335] R.A. Mah, M.R. Smith, and L. Baresi. Studies on an acetate-fermenting strain of methanosarcina. *Applied and Environmental Microbiology*, 35(6):1174–1184, 1978.

[336] M. P. Bryant and D. R. Boone. Emended description of strain MST(DSM 800t), the type strain of methanosarcina barkeri. *International Journal of Systematic Bacteriology*, 37(2):169–170, apr 1987.

[337] Maria V. Simankova, Sofja N. Parshina, Tatjana P. Tourova, Tatjana V. Kolganova, Alexander J.B. Zehnder, and Alla N. Nozhevnikova. Methanosarcina lacustris sp. nov., a new psychrotolerant methanogenic archaeon from anoxic lake sediments. *Systematic and Applied Microbiology*, 24(3):362–367, jan 2001.

[338] O. R. Kotsyurbenko, A. N. Nozhevnikova, T. I. Soloviova, and G. A. Zavarzin. Methanogenesis at low temperatures by microflora of tundra wetland soil. *Antonie van Leeuwenhoek*, 69(1):75–86, jan 1996.

[339] P.E. Rouviere and R.S. Wolfe. Observation of red fluorescence in a new marine methanosarcina. *88th Annual Meeting of the American Society of Microbiology*, page 184, 1988.

[340] M. A. Elberson and K. R. Sowers. Isolation of an aceticlastic strain of methanosarcina siciliae from marine canyon sediments and emendation of the species description for methanosarcina siciliae. *International Journal of Systematic Bacteriology*, 47(4):1258–1261, oct 1997.

[341] S. Ni and D. R. Boone. Isolation and characterization of a dimethyl sulfide-degrading methanogen, methanolobus siciliae HI350, from an oil well, characterization of m. siciliae t4/MT, and emendation of m. siciliae. *International Journal of Systematic Bacteriology*, 41(3):410–416, jul 1991.

[342] S. Shimizu, R. Upadhye, Y. Ishijima, and T. Naganuma. Methanosarcina horonobensis sp. nov., a methanogenic archaeon isolated from a deep subsurface miocene formation. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 61(10):2503–2507, nov 2010.

[343] Y. Liu, D.R. Boone, R. Sleat, and R.H. Mah. Methanosarcina-mazei lyc, a new methanogenic isolate which produces a disaggregating enzyme. *Applied and Environmental Microbiology*, 49(3):608–613, mar 1985.

[344] K.H. Blotevogel, U. Fischer, and K.H. Lupkes. Methanococcus-frisius sp-nov, a new methylotrophic marine methanogen. *Canadian Journal of Microbiology*, 32(2):127–131, Feb 1986.

[345] Susumu Asakawa, Masayo Akagawa-Matsushita, Hiroyuki Morii, Yosuke Koga, and Koichi Hayano. Characterization of methanosarcina mazeii TMA isolated from a paddy field soil. *Current Microbiology*, 31(1):34–38, jul 1995.

[346] U. Deppenmeier, M. Blaut, A. Jussofie, and G. Gottschalk. A methyl-CoM methylreductase system from methanogenic bacterium strain göl not requiring ATP for activity. *FEBS Letters*, 241(1-2):60–64, dec 1988.

[347] C. Joulian, B. Ollivier, B.K.C. Patel, and P.A. Roger. Phenotypic and phylogenetic characterization of dominant culturable methanogens isolated from ricefield soils. *FEMS Microbiology Ecology*, 25(2):135–145, feb 1998.

[348] T. D. Otto, M. Sanders, M. Berriman, and C. Newbold. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, 26(14):1704–1707, jun 2010.

[349] G. Srinivasan. Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA. *Science*, 296(5572):1459–1462, may 2002.

[350] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[351] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[352] Feng Chen, Aaron J. Mackey, Christian J. Stoeckert, and David S. Roos. Orthomcl-db: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34(suppl 1):D363–D368, 2006.

[353] S. van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, Amsterdam, Netherlands, May 2000.

[354] Stijn van Dongen and Cei Abreu-Goodger. *Using MCL to Extract Clusters from Networks*, chapter 15, pages 281–295. Springer New York, New York, NY, 2012.

[355] Kazutaka Katoh and Daron M. Standley. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[356] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696–704, Oct 2003.

[357] W. Hordijk and O. Gascuel. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, 21(24):4338–4347, oct 2005.

[358] E. Michael Gertz, Yi-Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. Composition-based statistics and translated nucleotide searches: Improving the tblastn module of blast. *BMC Biology*, 4(1):41, 2006.

[359] A.M Feist, M.J. Herrgard, I. Thiele, J.L. Reed, and B.Ø. Palsson. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology*, 7(2):129–143, 2009.

[360] Robert H White. Biosynthesis of the methanogenic cofactors. In *Vitamins & Hormones*, pages 299–337. Elsevier BV, 2001.

[361] J. L. Reed, T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong, and B. O. Palsson. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences*, 103(46):17480–17484, nov 2006.

[362] Vinay Satish Kumar and Costas D. Maranas. GrowMatch: An automated method for reconciling in silico/in vivo growth predictions. *PLoS Computational Biology*, 5(3):e1000308, mar 2009.

[363] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, mar 2003.

[364] Philip E. Luton, Jonathan M. Wayne, Richard J. Sharp, and Paul W. Riley. The mcrA gene as an alternative to 16s rRNA in the phylogenetic analysis of methanogen populations in landfill b. *Microbiology*, 148(11):3521–3530, nov 2002.

[365] Guillaume Borrel, Paul W. O'Toole, Hugh M.B. Harris, Pierre Peyret, Jean-François Brugère, and Simonetta Gribaldo. Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis. *Genome Biology and Evolution*, 5(10):1769–1780, aug 2013.

[366] Guillaume Borrel, Nicolas Parisot, Hugh MB Harris, Eric Peyretaillade, Nadia Gaci, William Tottey, Olivier Bardot, Kasie Raymann, Simonetta Gribaldo, Pierre Peyret, Paul W O'Toole, and Jean-François Brugère. Comparative genomics highlights the unique biology of methanomassiliicoccales, a thermoplasmatales-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics*, 15(1):679, 2014.

[367] Zhe Lyu and Yahai Lu. Comparative genomics of threeMethanocellalesstrains reveal novel taxonomic and metabolic features. *Environmental Microbiology Reports*, 7(3):526–537, apr 2015.

[368] Gonzalo Torres Tejerizo, Yong Sung Kim, Irena Maus, Daniel Wibberg, Anika Winkler, Sandra Off, Alfred Pühler, Paul Scherer, and Andreas Schlüter. Genome sequence of methanobacterium congolense strain buetzberg, a hydrogenotrophic, methanogenic archaeon, isolated from a mesophilic industrial-scale biogas plant utilizing bio-waste. *Journal of Biotechnology*, 247:1–5, apr 2017.

[369] Zhiliang Yu, Yunting Ma, Weihong Zhong, Juanping Qiu, and Jun Li. Comparative genomics of methanopyrus sp. SNP6 and KOL6 revealing genomic regions of plasticity implicated in extremely thermophilic profiles. *Frontiers in Microbiology*, 8, jul 2017.

[370] E. E. Hansen, C. A. Lozupone, F. E. Rey, M. Wu, J. L. Guruge, A. Narra, J. Goodfellow, J. R. Zaneveld, D. T. McDonald, J. A. Goodrich, A. C. Heath, R. Knight, and J. I. Gordon. Pan-genome of the dominant human gut-associated archaeon, methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences*, 108(Supplement_1):4599–4606, feb 2011.

[371] U. Deppenmeier, A. Johann, T. Hartsch, R. Merkl, R. A. Schmitz, R. Martinez-Arias, A. Henne, A. Wiezer, S. Baumer, C. Jacobi, H. Bruggemann, T. Lienard, A. Christmann, M. Bomeke, S. Steckel, A. Bhattacharyya, A. Lykidis, R. Overbeek, H. P. Klenk, R. P. Gunsalus, H. J. Fritz, and G. Gottschalk. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. *Journal of Molecular Microbiology and Biotechnology*, 4(4):453–461, Jul 2002.

[372] Emanuele Bosi, Jonathan M. Monk, Ramy K. Aziz, Marco Fondi, Victor Nizet, and Bernhard Ø. Palsson. Comparative genome-scale modelling of *staphylococcus aureus* strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proceedings of the National Academy of Sciences*, 113(26):E3801–E3809, jun 2016.

[373] Oksana Lukjancenko, Trudy M. Wassenaar, and David W. Ussery. Comparison of 61 sequenced escherichia coli genomes. *Microbial Ecology*, 60(4):708–720, jul 2010.

[374] Claudio Donati, N Luisa Hiller, Hervé Tettelin, Alessandro Muzzi, Nicholas J Croucher, Samuel V Angiuoli, Marco Oggioni, Julie C Dunning Hotopp, Fen Z Hu, David R Riley, Antonello Covacci, Tim J Mitchell, Stephen D Bentley, Morgens Kilian, Garth D Ehrlich, Rino Rappuoli, E Richard Moxon, and Vega Masignani. Structure and dynamics of the pan-genome of streptococcus pneumoniae and closely related species. *Genome Biology*, 11(10):R107, 2010.

[375] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, oct 2015.

[376] Y. Koga, M. Nishihara, H. Morii, and M. Akagawa-Matsushita. Ether polar lipids of methanogenic bacteria: structures, comparative aspects, and biosyntheses. *Microbiol. Rev.*, 57(1):164–182, Mar 1993.

[377] Benjamin H Meyer and Sonja-Verena Albers. *Archaeal Cell Walls*. John Wiley & Sons, Ltd, 2001.

[378] Katrin Weidenbach, Lisa Nickel, Horst Neve, Omer S. Alkhnbashi, Sven Künzel, Anne Kupczok, Thorsten Bauersachs, Liam Cassidy, Andreas Tholey, Rolf Backofen, and Ruth A. Schmitz. Methanosarcina spherical virus, a novel archaeal lytic virus targeting methanosarcina strains. *Journal of Virology*, 91(22):e00955–17, sep 2017.

[379] Benjamin J. Rauch, John Klimek, Larry David, and John J. Perona. Persulfide formation mediates cysteine and homocysteine biosynthesis in methanosarcina acetivorans. *Biochemistry*, 56(8):1051–1061, feb 2017.

[380] Scott I. Hauenstein and John J. Perona. Redundant synthesis of cysteinyl-tRNACysinMethanosarcina mazei. *Journal of Biological Chemistry*, 283(32):22007–22017, jun 2008.

[381] Y. Liu, A. Nakamura, Y. Nakazawa, N. Asano, K. A. Ford, M. J. Hohn, I. Tanaka, M. Yao, and D. Soll. Ancient translation factor is essential for tRNA-dependent cysteine biosynthesis in methanogenic archaea. *Proceedings of the National Academy of Sciences*, 111(29):10520–10525, jul 2014.

[382] Joe H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, mar 1963.

[383] T C Tallant and J A Krzycki. Methylthiol:coenzyme m methyltransferase from methanosarcina barkeri, an enzyme of methanogenesis from dimethylsulfide and methylmercaptopropionate. *Journal of Bacteriology*, 179(22):6902–6911, nov 1997.

[384] Thomas C. Tallant, Ligi Paul, and Joseph A. Krzycki. The mtsa subunit of the methylthiol:coenzyme m methyltransferase of methanosarcina barkeri catalyses both half-reactions of corrinoid-dependent dimethylsulfide: Coenzyme m methyl transfer. *Journal of Biological Chemistry*, 276(6):4485–4493, 2001.

[385] R. Jasso-Chavez, E. E. Apolinario, K. R. Sowers, and J. G. Ferry. MrpA functions in energy conversion during acetate-dependent growth of

methanosarcina acetivorans. *Journal of Bacteriology*, 195(17):3987–3994, jul 2013.

[386] Yan-Huai R Ding, Shi-Ping Zhang, Jean-Francois Tomb, and James G Ferry. Genomic and proteomic analyses reveal multiple homologs of genes encoding enzymes of the methanol:coenzyme m methyltransferase system that are differentially expressed in methanol- and acetate-grown methanosarcina thermophila. *FEMS Microbiology Letters*, 215(1):127–132, 2002.

[387] G. Kulkarni, D. M. Kridelbaugh, A. M. Guss, and W. W. Metcalf. Hydrogen is a preferred intermediate in the energy-conserving electron transport chain of methanosarcina barkeri. *Proceedings of the National Academy of Sciences*, 106(37):15915–15920, sep 2009.

[388] A. Jacobi, R. Rossmann, and A. Böck. The hyp operon gene products are required for the maturation of catalytically active hydrogenase isoenzymes in escherichia coli. *Archives of Microbiology*, 158(6):444–451, Nov 1992.

[389] Martin Vaupel and R. K. Thauer. Two f420-reducing hydrogenases in methanosarcina barkeri. *Archives of Microbiology*, 169(3):201–205, feb 1998.

[390] James G. Ferry. *Methanogenesis: Ecology, Physiology, Biochemistry & Genetics (Chapman & Hall Microbiology Series)*. Springer, 1994.

[391] T. Kupke. 4'-phosphopantetheine biosynthesis in archaea. *Journal of Biological Chemistry*, 281(9):5435–5444, dec 2005.

[392] Y. Liu, M. Sieprawska-Lupa, W. B. Whitman, and R. H. White. Cysteine is not the sulfur source for iron-sulfur cluster and methionine biosynthesis in the methanogenic archaeon methanococcus maripaludis. *Journal of Biological Chemistry*, 285(42):31923–31929, aug 2010.

[393] Yuchen Liu, Laura L. Beer, and William B. Whitman. Sulfur metabolism in archaea reveals novel processes. *Environmental Microbiology*, 14(10):2632–2644, may 2012.

[394] L. M. Proctor, R. Lai, and R. P. Gunsalus. The methanogenic archaeon Methanosarcina thermophila TM-1 possesses a high-affinity glycine betaine transporter involved in osmotic adaptation. *Appl. Environ. Microbiol.*, 63(6):2252–2257, Jun 1997.

[395] Marion KARRASCH, Gerhard BORNER, Marion ENSSLE, and Rudolf K. THAUER. The molybdoenzyme formylmethanofuran dehydrogenase from methanosarcina barkeri contains a pterin cofactor. *European Journal of Biochemistry*, 194(2):367–372, dec 1990.

[396] M. M. Wuebbens and K. V. Rajagopalan. Investigation of the early steps of molybdopterin biosynthesis in escherichia coli through the use of in vivo labeling studies. *Journal of Biological Chemistry*, 270(3):1082–1087, jan 1995.

[397] Christoph Rieder, Wolfgang Eisenreich, John O'Brien, Gerald Richter, Eva Gotze, Peter Boyle, Sylvain Blanchard, Adelbert Bacher, and Helmut Simon. Rearrangement reactions in the biosynthesis of molybdopterin. an NMR study with multiply 13c/15n labelled precursors. *European Journal of Biochemistry*, 255(1):24–36, jul 1998.

[398] M. M. Wuebbens and K. V. Rajagopalan. Mechanistic and mutational studies of escherichia coli molybdopterin synthase clarify the final step of molybdopterin biosynthesis. *Journal of Biological Chemistry*, 278(16):14523–14532, feb 2003.

[399] R. Thome, A. Gust, R. Toci, R. Mendel, F. Bittner, A. Magalon, and A. Walburger. A sulfurtransferase is essential for activity of formate dehydrogenases in escherichia coli. *Journal of Biological Chemistry*, 287(7):4671–4678, dec 2011.

[400] Pascal Arnoux, Christian Ruppelt, Flore Oudouhou, Jérôme Lavergne, Marina I. Siponen, René Toci, Ralf R. Mendel, Florian Bittner, David Pignol, Axel Magalon, and Anne Walburger. Sulphur shuttling across a chaperone during molybdenum cofactor maturation. *Nature Communications*, 6:6148, feb 2015.

[401] Loes E. Bevers, Peter-Leon Hagedoorn, José A. Santamaria-Araujo, Axel Magalon, Wilfred R. Hagen, and Guenter Schwarz. Function of MoaB proteins in the biosynthesis of the molybdenum and tungsten cofactors†. *Biochemistry*, 47(3):949–956, jan 2008.

[402] A. Llamas, T. Otte, G. Multhaup, R. R. Mendel, and G. Schwarz. The mechanism of nucleotide-assisted molybdenum insertion into molybdopterin: A NOVEL ROUTE TOWARD METAL COFACTOR ASSEMBLY. *Journal of Biological Chemistry*, 281(27):18343–18350, apr 2006.

[403] Kaiyuan Zheng, Phong D. Ngo, Victoria L. Owens, Xue peng Yang, and Steven O. Mansoorabadi. The biosynthetic pathway of coenzyme f430 in methanogenic and methanotrophic archaea. *Science*, 354(6310):339–342, oct 2016.

[404] Simon J. Moore, Sven T. Sowa, Christopher Schuchardt, Evelyne Deery, Andrew D. Lawrence, José Vazquez Ramos, Susan Billig, Claudia Birkemeyer, Peter T. Chivers, Mark J. Howard, Stephen E. J. Rigby, Gunhild Layer, and Martin J. Warren. Elucidation of the biosynthesis of the methane catalyst coenzyme f430. *Nature*, feb 2017.

[405] D. Möller-Zinkhan, G. Börner, and R. K. Thauer. Function of methanofuran, tetrahydromethanopterin, and coenzyme f420 in archaeoglobus fulgidus. *Archives of Microbiology*, 152(4):362–368, sep 1989.

[406] J. A. Vorholt, Doris Hafenbradl, Karl O. Stetter, and Rudolf K. Thauer. Pathways of autotrophic CO 2 fixation and of dissimilatory nitrate reduction to n 2 o in ferroglobus placidus. *Archives of Microbiology*, 167(1):19–23, jan 1997.

[407] W. Hordijk and O. Gascuel. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, 21(24):4338–4347, oct 2005.

[408] Wulf Blankenfeldt and James F Parsons. The structural biology of phenazine biosynthesis. *Current Opinion in Structural Biology*, 29:26–33, dec 2014.

[409] Zhen Yan, Mingyu Wang, and James G. Ferry. A ferredoxin- and f420h2-dependent, electron-bifurcating, heterodisulfide reductase with homologs in the domains bacteria and archaea. *mBio*, 8(1):e02285–16, feb 2017.

[410] S. Wang, J. Tiongson, and M. E. Rasche. Discovery and characterization of the first archaeal dihydromethanopterin reductase, an iron-sulfur flavoprotein from methanosarcina mazei. *Journal of Bacteriology*, 196(2):203–209, aug 2013.

[411] D. E. McNamara, D. Cascio, J. Jorda, C. Bustos, T.-C. Wang, M. E. Rasche, T. O. Yeates, and T. A. Bobik. Structure of dihydromethanopterin reductase, a cubic protein cage for redox transfer. *Journal of Biological Chemistry*, 289(13):8852–8864, feb 2014.

[412] J. J. Zulty and M. K. Speedie. Purification and characterization of S-adenosylhomocysteine deaminase from streptonigrin-producing Streptomyces flocculus. *J. Bacteriol.*, 171(12):6840–6844, Dec 1989.

[413] Samta Jain, Antonella Caforio, Peter Fodran, Juke S. Lolkema, Adriaan J. Minnaard, and Arnold J.M. Driessen. Identification of CDP-archaeol synthase, a missing link of ether lipid biosynthesis in archaea. *Chemistry & Biology*, 21(10):1392–1401, oct 2014.

[414] Yihong Chen, Ethel Apolinario, Libuse Brachova, Zvi Kelman, Zhuo Li, Basil J Nikolau, Lucas Showman, Kevin Sowers, and John Orban. A nuclear magnetic resonance based approach to accurate functional annotation of putative enzymes in the methanogen methanosarcina acetivorans. *BMC Genomics*, 12(Suppl 1):S7, 2011.

[415] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8614–8619, jul 2001.

[416] Ertugrul M. Ozbudak, Mukund Thattai, Iren Kurtser, Alan D. Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, apr 2002.

[417] Mads Kærn, Timothy C. Elston, William J. Blake, and James J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, may 2005.

[418] D. Schultz, E. Ben Jacob, J. N. Onuchic, and P. G. Wolynes. Molecular level stochastic model for competence cycles in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17582–17587, oct 2007.

[419] P. J. Choi, L. Cai, K. Frieda, and X. S. Xie. A Stochastic Single-Molecule Event Triggers Phenotype Switching of a Bacterial Cell. *Science*, 322(5900):442–446, oct 2008.

[420] Murat Acar, Jerome T Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475, 2008.

[421] Mingyang Lu, José Onuchic, and Eshel Ben-Jacob. Construction of an effective landscape for multistate genetic switches. *Physical Review Letters*, 113(7), aug 2014.

[422] L. T. MacNeil and A. J. M. Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, 21(5):645–657, feb 2011.

[423] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45):17256–17261, nov 2008.

[424] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[425] Lok So, Anandamohan Ghosh, Chenghang Zong, Leonardo A Sepúlveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in *Escherichia coli*. *Nature Genetics*, 43(6):554–560, may 2011.

[426] J. Peccoud and B. Ycart. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, oct 1995.

[427] Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036, dec 2005.

[428] Hinrich Boeger, Joachim Griesenbeck, and Roger D. Kornberg. Nucleosome Retention and the Stochastic Nature of Promoter Chromatin Remodeling for Transcription. *Cell*, 133(4):716–726, may 2008.

[429] Michael J. Hallock, John E. Stone, Elijah Roberts, Corey Fry, and Zaida Luthey-Schulten. Simulation of reaction diffusion processes over biologically relevant size and time scales using multi-GPU workstations. *Parallel Computing*, 40(5-6):86–99, may 2014.

[430] S. Cooper and C. E. Helmstetter. Chromosome replication and the division cycle of Escherichia coli B/r. *Journal of Molecular Biology*, 31(3):519–540, Feb 1968.

[431] H. E. Kubitschek and M. L. Freedman. Chromosome replication and the division cycle of Escherichia coli B-r. *Journal of Bacteriology*, 107(1):95–99, Jul 1971.

[432] Po-Yi Ho and Ariel Amir. Simultaneous regulation of cell size and chromosome replication in bacteria. *Frontiers in Microbiology*, 6, jul 2015.

[433] K. Skarstad, E. Boye, and H. B. Steen. Timing of initiation of chromosome replication in individual *Escherichia coli* cells. *EMBO Journal*, 5(7):1711–1717, Jul 1986.

[434] EO Powell. Growth rate and generation time of bacteria, with special reference to continuous culture. *Journal of General Microbiology*, 15(3):492–511, 1956.

[435] K. Kimata, T. Inada, H. Tagami, and H. Aiba. A global repressor (Mlc) is involved in glucose induction of the ptsG gene encoding major glucose transporter in *Escherichia coli*. *Molecular Microbiology*, 29(6):1509–1519, Sep 1998.

[436] S. Klumpp and T. Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20245–20250, dec 2008.

[437] U. Vogel and K. F. Jensen. The RNA chain elongation rate in Escherichia coli depends on the growth rate. *Journal of Bacteriology*, 176(10):2807–2813, May 1994.

[438] G Reshes, S Vanounou, I Fishov, and M Feingold. Timing the start of division in *E. coli*: a single-cell study. *Physical Biology*, 5(4):046001, nov 2008.

[439] S. El Qaidi and J. Plumbridge. Switching Control of Expression of ptsG from the Mlc Regulon to the NagC Regulon. *Journal of Bacteriology*, 190(13):4677–4686, may 2008.

[440] P. B. Eckburg. Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–1638, jun 2005.

[441] Kathryn J. Pflughoeft and James Versalovic. Human microbiome in health and disease. *Annual Review of Pathology: Mechanisms of Disease*, 7(1):99–122, feb 2012.

[442] Kevin R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355, sep 2005.

[443] Marcel G. A. van der Heijden, Richard D. Bardgett, and Nico M. van Straalen. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology Letters*, 11(3):296–310, mar 2008.

[444] Peter J. Turnbaugh, Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, oct 2007.

[445] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E Sands, Ramnik J Xavier, and Curtis Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012.

[446] Dirk Gevers, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, Dan Knights, Se Jin Song, Moran Yassour, Xochitl C. Morgan, Aleksandar D. Kostic, Chengwei Luo, Antonio González, Daniel McDonald, Yael Haberman, Thomas Walters, Susan Baker, Joel Rosh, Michael Stephens, Melvin Heyman, James Markowitz, Robert Baldassano, Anne Griffiths, Francisco Sylvester, David Mack, Sandra Kim, Wallace Crandall, Jeffrey Hyams, Curtis Huttenhower, Rob Knight, and Ramnik J. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host & Microbe*, 15(3):382–392, mar 2014.

[447] Hsuan-Chao Chiu, Roie Levy, and Elhanan Borenstein. Emergent biosynthetic capacity in simple microbial communities. *PLoS Computational Biology*, 10(7):e1003695, jul 2014.

[448] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, nov 2009.

[449] Michael A. Henson and Timothy J. Hanly. Dynamic flux balance analysis for synthetic microbial communities. *IET Systems Biology*, 8(5):214–229, oct 2014.

[450] Matthew B. Biggs, Gregory L. Medlock, Glynis L. Kolling, and Jason A. Papin. Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334, jun 2015.

[451] Hank Childs, Eric Brugger, Brad Whitlock, Jeremy Meredith, Sean Ahern, David Pugmire, Kathleen Biagas, Mark Miller, Cyrus Harrison, Gunther H. Weber, Hari Krishnan, Thomas Fogal, Allen Sanderson, Christoph Garth, E. Wes Bethel, David Camp, Oliver Rübel, Marc Durant, Jean M. Favre, and Paul Navrátil. VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data. In *High Performance Visualization–Enabling Extreme-Scale Scientific Insight*, pages 357–372. Chapman & Hall/CRC, Oct 2012.

[452] John A. Cole and Zaida Luthey-Schulten. Whole cell modeling: From single cells to colonies. *Israel Journal of Chemistry*, 54(8-9):1219–1229, jul 2014.

[453] William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Amrita Kar, Nicholas Leiby, Pankaj Mehta, Christopher J. Marx, and Daniel Segrè. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7(4):1104–1115, may 2014.

[454] Jin Chen, Jose A. Gomez, Kai Höffner, Poonam Phalak, Paul I. Barton, and Michael A. Henson. Spatiotemporal modeling of microbial metabolism. *BMC Systems Biology*, 10(1), March 2016.

[455] Poonam Phalak, Jin Chen, Ross P. Carlson, and Michael A. Henson. Metabolic modeling of a chronic wound biofilm consortium predicts spatial partitioning of bacterial species. *BMC Systems Biology*, 10(1), sep 2016.

[456] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. A comprehensive genome-scale reconstruction of escherichia coli metabolism–2011. *Molecular Systems Biology*, 7(1):535–535, apr 2014.

[457] Jonathan M. Monk, Anna Koza, Miguel A. Campodonico, Daniel Machado, Jose Miguel Seoane, Bernhard O. Palsson, Markus J. Herrgård, and Adam M. Feist. Multi-omics quantification of species variation of escherichia coli links molecular features with strain phenotypes. *Cell Systems*, 3(3):238–251.e12, sep 2016.

[458] Zachary A. King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson,

and Nathan E. Lewis. BiGG models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44(D1):D515–D522, oct 2015.

[459] Jeffrey D. Orth, Bernhard Ø. Palsson, and R. M. T. Fleming. Reconstruction and use of microbial metabolic networks: the core escherichia coli metabolic model as an educational guide. *EcoSal Plus*, 4(1), sep 2010.

[460] Helena Mendes-Soares, Michael Mundy, Luis Mendes Soares, and Nicholas Chia. MMinte: an application for predicting metabolic interactions among the microbial species in a community. *BMC Bioinformatics*, 17(1), sep 2016.

[461] Milan J.A. van Hoek and Roeland M.H. Merks. Emergence of microbial diversity due to cross-feeding interactions in a spatial model of gut microbial metabolism. *BMC Systems Biology*, 2016.

[462] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, nov 2016.

[463] Areti Tsigkinopoulou, Syed Murtuza Baker, and Rainer Breitling. Respectful modeling: Addressing uncertainty in dynamic system models for molecular biology. *Trends in Biotechnology*, 35(6):518–529, jun 2017.